Characterisation of a *bagpipe* homologue in *Xenopus laevis*

A thesis by:

Nicola D. Civill

Submitted to the University of London in fulfilment of the requirements for the degree: Doctor of Philosophy

September 1999

Supervisors:

Dr T.J. Mohun

Division of Developmental Biology

National Institute for Medical Research

The Ridgeway

Mill Hill

London

NW7 1AA

Dr Nugent

Department of Biology

University College London

Gower Street

London

WC1E 6BT

God be with those who explore in the cause of understanding, whose search takes them far from what is familiar and comfortable and leads them into danger or terrifying loneliness. Let us try to understand their sometimes strange or difficult ways; their confronting or unusual language; the uncommon life of their emotions, for they have been affected and shaped by their struggle at the frontiers of a wild darkness, just as we may be affected, shaped and changed by the insight they bring back to us.

M. Leunig

For Paul Norris, Maureen Hibbs and little Cara Jayne

Acknowledgments

This thesis was funded by the Medical Research Council and was conducted at the National Institute for Medical research under the Directorship of Sir John Skehel in the Department of Developmental Biology under Dr. Jim Smith all of whom I am very grateful to.

There are many people that deserve thanks for their contribution to the completion of this thesis. Firstly, my supervisor Dr. Tim Mohun for his technical and theoretical guidance through the long journey of discovery and for his support of my chosen path beyond NIMR. I am indebted to the 'Cloning Queen' Norma Towers without whose pearls of wisdom cloning would have been a far greater trial. Also, to the 'Sequencing King' Surendra Kotecha for his help with the binding site selection. Jacky Smith made my time at NIMR much easier with great technical support and excellent theatre tips! The entire department has provided me with much encouragement and scientific discussion over the last three years, with particular thanks to Dr. Elena Casey for wading through the first versions of my thesis. Brian Cooper deserves a special mention for his renditions of the Adams Family, The Muppet Show and many more tunes that have kept me laughing the last couple of years! The photographics department at NIMR has made the final production of this thesis far less painful. Many thanks also go to Mark Bradburn from the Department of Mathematical Biology for conducting the phylogenetic analysis.

The student network has been invaluable for cross-discipline technical tips and emotional support. I am particularly grateful to fellow students Karen Allman, Ruwani Gunaratne, Nina Collins and Taslima Khan, for always being there for me. I must sincerely thank my husband Paul Norris for his continuing support and for supplying chocolate and cookies to help me through the most stressful periods, and finally to my Mum, Maureen Hibbs, for always believing in me.

Abstract

Mutations of *Drosophila bagpipe*, an *NK3* class homeobox gene, result in failure of visceral mesoderm to differentiate into stomach tissue. Thus *bagpipe*, in association with other factors, is a good candidate for specification of visceral mesoderm in *Drosophila*. The cloning and characterisation of a *Xenopus bagpipe* homologue was therefore of great interest. This study describes the isolation of a full length *Xenopus* cDNA clone that on the basis of database analysis and sequence comparisons has been assigned as *Xenopus bagpipe* (*XBap*).

Previous studies had revealed that the majority of homeoproteins recognise DNA sites with a 5'-TAAT-3' core but the NK class of homeoproteins had been shown to bind specifically to sites containing a 5'-CAAG-3' core. Experiments described here, however, show the XBap DNA binding site to be an even more divergent, 5'-TTAAGTGG---TTAAGTGG-3'. A series of mutant oligonucleotides revealed that the 'T' of the 5'-$T_1A_2A_3G_4$-3' core, as well as the presence of two such cores, are indeed essential for optimal XBap DNA binding. The murine NK3 class homeoproteins, Nkx-3.1 and Bapx1, are demonstrated to have the same requirements for optimal DNA binding as XBap. *Drosophila* Bagpipe, however, was found to have a less stringent requirement for a 'T' at position one of the core, binding equally well to a 'C' in this position, but the presence of two such cores is still necessary for optimal DNA binding. Preliminary studies using site directed mutagenesis attempted to define the amino acids responsible for the differences.

The effect of XBap on transcription was studied using a *Xenopus* oocyte assay and two cell transfection assays. XBap was not found to act as a transcriptional activator in any of these assays but evidence was obtained to suggest that a C-terminal truncation of XBap could act as a repressor of transcription.

Abbreviations

AA	amino acid
ATP	adenosine-5'triphosphate
BSA	bovine serum albumin
CAT	chloramphenicol acetyl transferase
C-terminal	carboxy-terminal
Ci	curie
CsCl	caesium chloride
CTAB	hexadecyltrimethylammonium bromide
DMut	double-mutant
DTT	dithiothreitol
E.Coli	*Escherischia coli*
EDTA	diaminoethanetetra-acetic acid, disodium salt
EGTA	ethyleneglycol-bis-(b-aminoethyl ether) N, N, N', N'-tetra-acetic acid
EMSA	Electrophoretic Mobility Shift Assay
GTP	guanosine-5'-triphosphate
HCl	hydrochloric acid
HD	homeodomain
IgG	Immunoglobulin G
Kb	kilobase
Kd	kilodalton
KCl	potassium chloride
Klenow	Klenow fragment of DNA polymerase I
LiCl	lithium chloride
MMLV	Moloney mouse leukaemia virus
MOPS	(3-[N-morpholino]propanesulphonic acid)
$MgCl_2$	magnesium chloride
$MgSO_4$	magnesium sulphate

N-terminal	amino-terminal
NaCl	sodium chloride
NAM	Normal amphibian medium
NH_4Ac	ammonium acetate
NIMR	National Institute for Medical Research
1X PCR Buffer	50mM KCl, 10mM Tris-HCl, 2.5mM $MgCl_2$, 0.2mM dATP, dGTP, dCTP and dTTP, 0.17mg/ml BSA
ORT	(NaCl 82.5mM, KCl 2.5 mM, Na_2HPO_4 1 mM, HEPES 5 mM, PVP 0.5g/l)
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PIPES	piperazine-N, N'-bis[2-etanesuphonic acid]
PVP	polyvinylpyrrolidone
RNA	ribonucleic acid
RNase	ribonuclease
RT-PCR	reverse transcriptase- polymerase chain reaction
RT	room temperature
SDS	sodium dodecyl sulphate
SMut	single-mutant
SSC	1X= 150mM sodium chloride, 15mM sodium citrate, pH 7.5
TLC	Thin layer chromatography
Trizma	tris[hydroxymethyl]aminomethane
TTP	thymidine triphosphate
UTR	untranslated region
UTP	uridine triphosphate
bp	base pair
cDNA	complementary deoxyribonucleic acid
dNTPs	deoxyribonucleoside triphosphates
MBq	Mega bequerels

ml	millilitre
mM	millimolar
µg	microgram
µl	microlitre

Table of Figures

Chapter One

General Introduction

1.0 Chapter 1: General Introduction

1.1 Homeobox genes

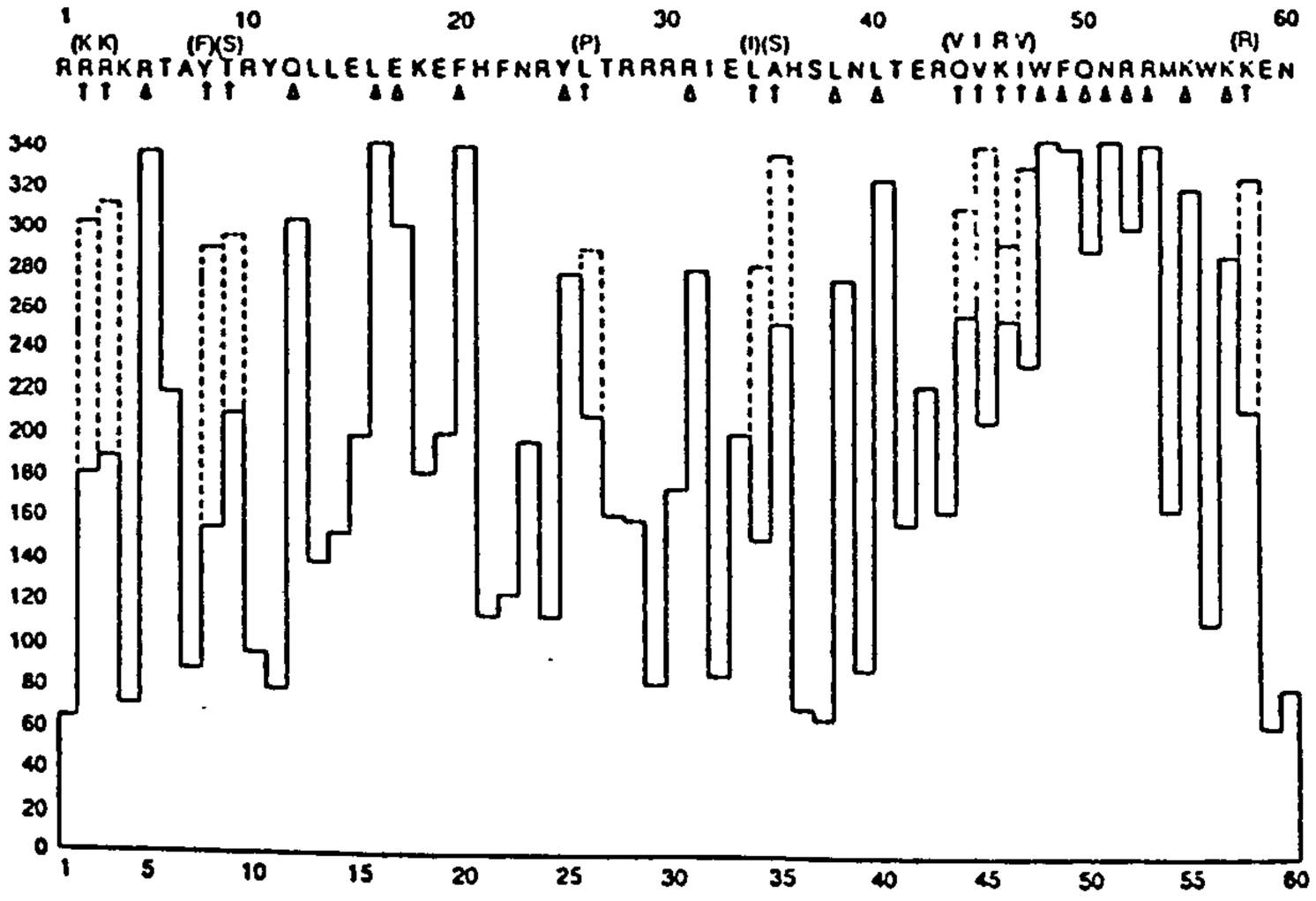
The homeodomain was originally described as a highly conserved protein motif of 60 amino acids, encoded by a short DNA fragment, the homeobox, that was found in *Drosophila melanogaster* developmental control genes (McGinnis *et al*, 1994). Homeobox containing genes have since been identified in every eukaryote investigated, spanning a broad evolutionary spectrum from yeast to man. The homeodomain represents the DNA-binding domain of much larger proteins, which allows the sequence-specific recognition of target genes. These homeoproteins primarily serve a gene regulatory function, acting as transcription factors, which regulate genes in a precise temporal and spatial pattern (Gehring *et al*, 1994). Originally believed to be the key to segment identity in *Drosophila* and vertebrates, the homeobox has assumed a much broader role in recent years (Hoey and Levine, 1988).

1.2 Sequence analysis of the homeodomain

Thomas R. Burglin compiled a consensus sequence based on a compilation of 346 homeodomain sequences (Fig.1) (Burglin *et al*, 1994). There are seven positions within the consensus that are occupied by the same amino acid in more than 95% of the 346 sequences (black triangles), 10 positions which are conserved in more than 80% of the sequences (open triangles) and 12 positions at which only two different amino acids are observed in more than 80% of the sequences. These highly conserved amino acids are reflected in the highest pillars of the frequency diagram and serve to define the homeodomain (Fig.1). The similarity of frequency diagrams produced from the alignment of the homeodomains of many individual species indicates a high degree of evolutionary conservation (Gehring *et al*, 1994 A, Burglin *et al*, 1994).

Figure. 1 Consensus homeodomain amino acid sequence

The consensus sequence derived from 346 homeodomain sequences. In the frequency diagram below, the highest amino acid frequencies for positions 1-60 are given. For those positions at which two amino acids are found very frequently, the frequency of the second amino acid is indicated by a dashed line and the respective amino acid shown above the consensus in parentheses. (Figure reproduced from, Gehring *et al*, 1994 A).


1 10 20 30 40 50 60
(K K) (F)(S) (P) (I)(S) (V I R V) (R)
R R R K R T A Y T R Y Q L L E L E K E F H F N R Y L T R R R R I E L A H S L N L T E R Q V K I W F Q N R R M K W K K E N
340
320
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
1 5 10 15 20 25 30 35 40 45 50 55 60

1.3 Structural analysis of the homeodomain

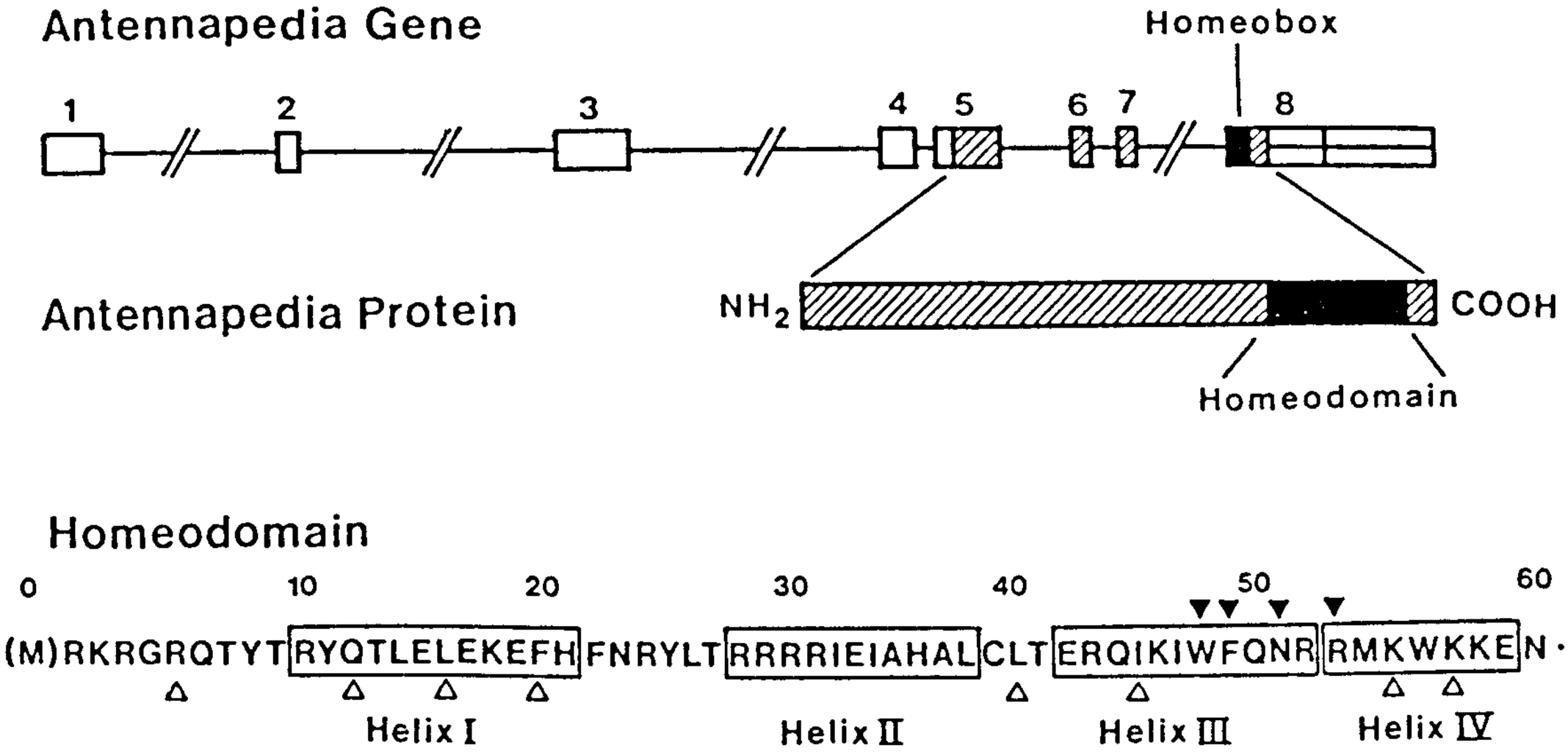
The molecular architecture of the homeodomain is held together by a core of 11 amino acids which are tightly packed. All but one of these residues are hydrophobic, and six (L16, L38, L40, I45, W48, F49) are very highly conserved or invariant amongst homeoproteins (Fig. 1). Since the hydrophobic core of a protein plays a key role in determining its structure, and so many of these core amino acids are highly conserved, one can predict that different homeodomains fold into similar three-dimensional structures. The discovery that the crystal structures of the Engrailed (En) and MATα2 homeodomains complexed with DNA clearly have similar molecular architectures to that determined for the free Antennapedia (Antp) in solution supports such a prediction (Gehring *et al*, 1994 A).

1.3.1 Structure of Antennapedia as determined by NMR spectroscopy

The solution structure of a 68 amino acid poly-peptide corresponding to the Antp homeodomain was determined at high resolution by NMR spectroscopy (Qian *et al*, 1988). The Antp homeodomain structure consists of helix I (residues 10-21), a connecting loop (residues 22-27) leading to helix II (residues 28-38), a tight turn of three residues (39-41), to helix III (residues 42-52), and a more disordered and flexible helix IV consisting of residues 53-59. The elongation of the third helix by the fourth helix is a structural element that so far appears to be unique to the Antp homeodomain. The two ends of the polypeptide chain (residues 0-6 and 60-67) are, in contrast, not well defined by the NMR spectra, presumably because they are flexible in solution (Qian *et al*, 1988, Gehring *et al*, 1994 B). The sequence of the Antp homeodomain with its corresponding structural features is illustrated in figure 2 (Gehring *et al*, 1990). Helices I and II are arranged in an anti-parallel fashion relative to each other, whereas helix III and its direct extension helix IV, are aligned essentially perpendicularly to the first two helices. Helices II and III form a helix-turn-helix motif as previously described for prokaryotic gene regulatory proteins (Gehring *et al*, 1994 A).

Figure 2. Structural organisation of the *Antennapedia* gene, the Antennapedia protein and its homeodomain.

Exons 1-8 are separated by introns, which are not drawn to scale. The homeobox is located in exon 8, which differs in length depending on which polyadenylation site is used. The locations of the four helices within the amino acid sequence of the homeodomain are indicated. The invariant amino acids are marked by black triangles, the highly conserved ones by open triangles (Figure reproduced from Gehring *et al*, 1990).


Antennapedia Gene
Homeobox
1
2
3
4
5
6
7
8
Antennapedia Protein
NH2
COOH
Homeodomain
Homeodomain
0
10
20
30
40
50
60
(M)RKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN·····
Helix I
Helix II
Helix III
Helix IV

The structure of a 1:1 complex between the Antp homeodomain and a 14bp DNA fragment in aqueous solution has also been determined by NMR (Otting *et al*, 1990). Residue 39 of Antp was changed from cysteine to serine to prevent oxidative dimerization of the protein (Antp39C-S). The DNA fragment corresponds to the BS2 site, , which had been identified as a specific binding site by DNA footprinting and gel retardation assays (BS2 $C_{-3}T_{-2}C_{-1}T_0T_1T_2T_3T_4C_5G_6G_7T_8A_9A_{10}T_{11}C_{12}T_{13}C_{14}$, Muller *et al*, 1988, also see Fig.3). This study revealed that the homeodomain interacts with both the major groove and the minor groove of the DNA, and establishes numerous contacts with the DNA backbone (Fig. 4) (Gehring *et al*, 1994 A).

Within the N-terminal arm two arginine residues contact the DNA in the minor groove. Arg3 forms a salt bridge to the phosphate group of G12 and Arg5 makes hydrophobic contacts to the sugar moieties of G12 and A13 as well as short contacts to the bases T11 and G12. Gln6 and Tyr8 contact the β-strand of the DNA via the backbone groups of A10 (Gehring *et al*, 1994 A).

Numerous contacts are observed between helices III/IV ('the recognition helix') and both strands of the DNA. Salt bridges from the side chains of Arg43, Arg52 and Lys55 to phosphate groups on the β-strand, and from Arg53 to the α-strand, connect the recognition helix to both sides of the major groove. Hydrophobic interactions from Arg43 and Ile47 to the β-strand and from Met54 and Lys46 to the α-strand provide additional backbone contacts further supplemented by contacts from Arg43, Gln44 and Met54 to phosphate groups.

Amino acid side chain contacts to DNA bases; Ile4 to T8 and A9, Gln50 to C7 and T8, and Met54 to C7 may provide sequence specific recognition. The loop between helices I and II and the start of helix II approaches the α-strand of the DNA. The side chain of Tyr25 contacts the sugar moiety of G5 and the phosphate of C6, while the side chains of

Figure 3. DNA binding sites used in structural and specificity studies

Antennapedia, Muller *et al*, 1988, Matα2, Wolberger *et al*, 1991, Engrailed, Kissenger *et al*, 1991, rat TTF-1, Guazzi *et al*, 1990.

Antennapedia 'BS2'		5'-$C_{-3}T_{-2}C_{-1}T_0T_1T_2T_3T_4C_5G_6G_7T_8A_9A_{10}T_{11}C_{12}T_{13}C_{14}$-3'
Matα2		5'-ACATGTAATTCATTTACACGC-3'
Engrailed		5'-TGCCATGTAATTACCTA –3'
TTF-1	'Oligo C'	5'-$C_{-2}A_{-1}G_0T_1C_2A_3A_4G_5T_6G_7T_8T_9$-3'

Figure 4. Schematic drawings of the Antp(C39S) homeodomain-DNA complex

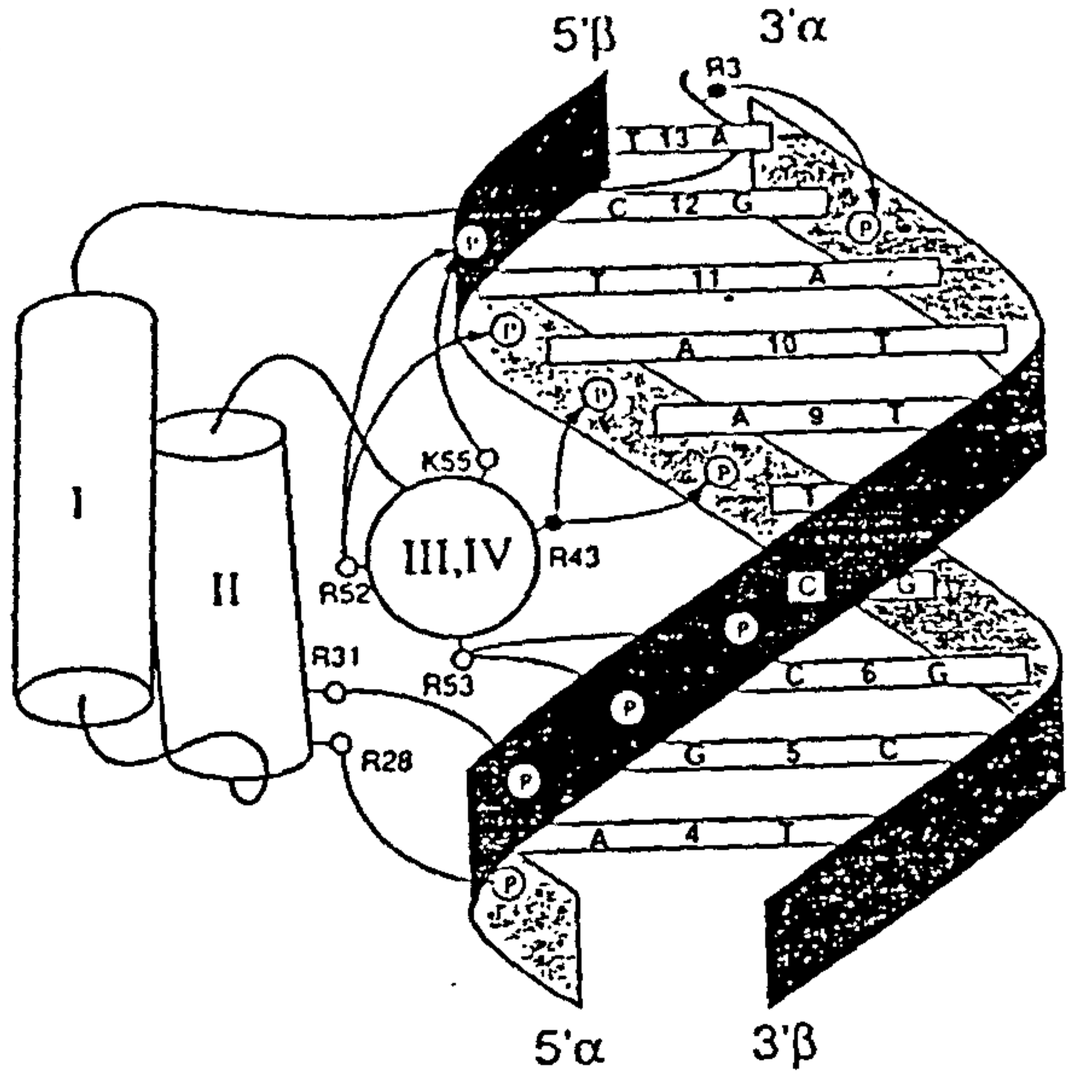
The view is along the axis of helix III from the C- to the N-terminus. Cylinders represent the helices of the homeodomain, the amino acid side chains that contact the DNA are identified with circles, where filled circles illustrate residues for which direct intermolecular contacts with the DNA have been identified. The DNA backbone is represented by ribbons and the DNA base pairs by horizontal bars with the single-letter symbols of the two bases. The four bars with the consensus base pairs 5'-TAAT-3' are shaded. Selected phosphate groups are identified by open circles with the letter 'P'. Arrows indicate short contacts between individual amino acid residues of the homeodomain and the DNA. Arrows ending within the bars indicate specific contacts with the DNA bases, those pointing to the open circles with the letter P indicate contacts to the phosphate groups and all other arrows indicate contacts with the deoxyribose moieties of the DNA.

A. Intermolecular electrostatic interactions with interatomic distances shorter than 5 angstroms.

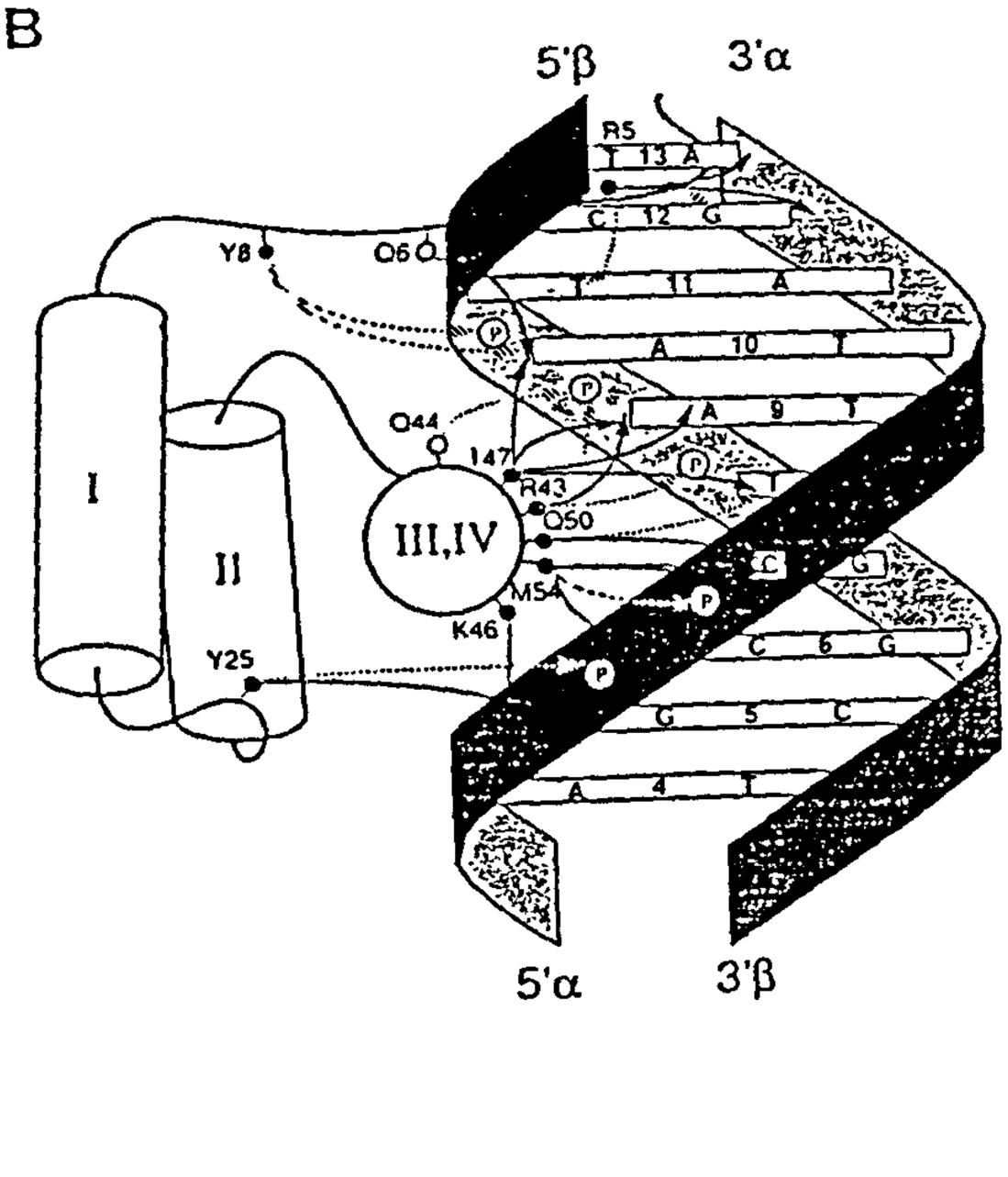
B. Hydrophobic contacts with interatomic distances shorter then 3.5 angstroms are shown with solid arrows, and other, miscellaneous contacts with interatomic distances shorter than 3.5 angstroms, including intermolecular hydrogen bonds are shown with dashed arrows.

(Figure reproduced from Gehring *et al*, 1994 B)

A


5'β
3'α
R3
T 13 A
C 12 G
T 11 A
A 10 T
A 9 T
T
C
G
C 6 G
G 5 C
A 4 T
P
K55
III,IV
R43
R52
R31
R53
R28
I
II
5'α
3'β

B


5'β
3'α
R5
T 13 A
C 12 G
T 11 A
A 10 T
A 9 T
T
C
G
C 6 G
G 5 C
A 4 T
P
Y8
O6
O44
I47
R43
O50
M54
K46
Y25
III,IV
I
II
5'α
3'β

Arg28 and Arg31 form salt bridges to the phosphate groups of A4 and G5, respectively. (Gehring *et al*, 1994 A).

The study also revealed that the Antp (C39-S) homeodomain-DNA complex contains a well-defined cavity at the interface between the recognition helix and the DNA duplex, which may contain up to five water molecules. It is possible that these water molecules make contacts across the cavity possible by mediating specific hydrogen bonds between Gln50 and Asn51 on the one hand and the polar groups of the DNA on the other.

1.3.2 X-ray crystallography studies of yeast MATα2 and *Dros* Engrailed

The MATα2 homeodomain regulates the expression of cell type-specific genes in yeast. The co-crystal structure of MATα2 homeodomain complexed with its biologically relevant operator DNA sequence, has been determined at 2.7 A resolution (Fig.3, 5'-ACATGTAATTCATTTACACGC-3' Wolberger *et al*, 1991). The co-crystal structure of *Drosophila* Engrailed (en) complexed to DNA has also been determined to 2.8A resolution, although it is not known whether the binding site used is functional *in vivo* (Fig.3, 5'-TGCCATGTAATTACCTA –3' Kissenger *et al*, 1991). The X-ray crystallographic data are in good agreement with the Antp NMR data with the overall polypeptide structures being quite similar despite the fact that Engrailed and MATα2 share only 27% sequence identity (Fig. 5) (Gehring *et al*, 1994 A).

The N-terminal arm fits into the minor groove in all three complexes with Arg3 and Arg5 of en, and Arg7 of Matα2 making contacts with bases in the DNA. Phe8 (Tyr8 in Antp) makes identical contacts to phosphates in the DNA backbone in all three complexes. Tyr 25 and Lys55 contacts dock the recognition helix into the major groove. Engrailed forms direct contacts with the bases in the major groove via residues Ile47, Gln50, and Asn51 of helix III although the side chain of Ala54 is too short to reach the DNA bases. In both the En and MATα2 complexes, the invariant Asn51 makes a pair of hydrogen bonds to an adenine in the TAAT core binding site, which in Antp are thought to be mediated by

Figure 5. Schematic comparison of the homeodomain-DNA complexes of Antp, En and Matα2

A. Antp. View perpendicular to the axis of the recognition helix (III and IV) located in the major groove. The α helices I, II and III, IV are indicated by cylinders. For the N-terminal arm (upper left), the loop between helices I and II, and the turn of the helix-turn-helix motif (II-III) only the polypeptide backbone is drawn as a solid line.

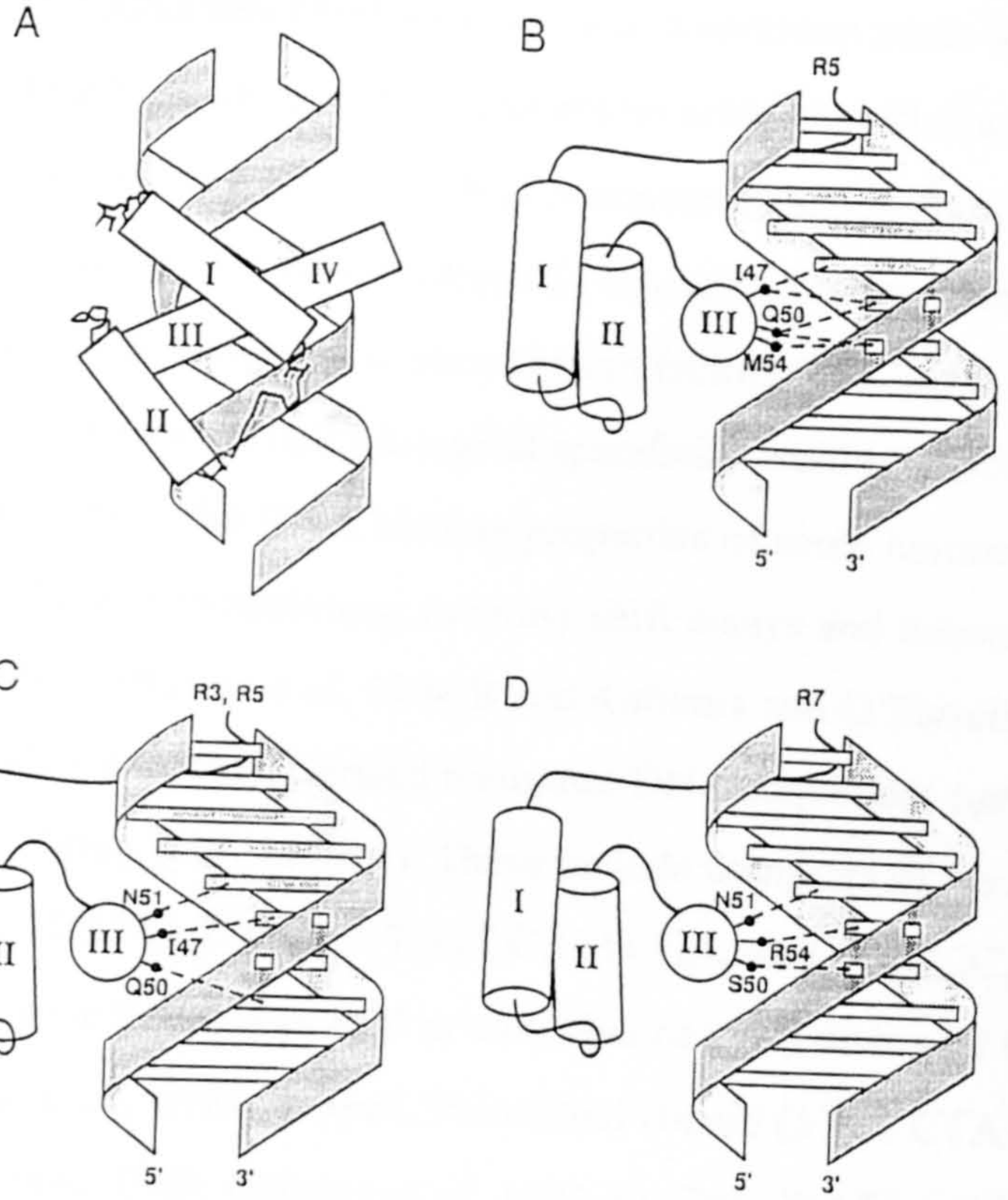
B. Antp

C. En

D. Matα2

View along the axis of helix III. Amino acid residues that establish contacts to specific bases are indicated (dotted lines). Residues in the recognition helix, which form contacts are in the major groove but residues in the N-terminal arm contact the minor groove of the DNA. In B. and C., the TAAT core motif is shaded.

(Figure reproduced from Gehring *et al*, 1994 B).


A
I
IV
III
II
B
R5
I
II
III
I47
Q50
M54
5'
3'
C
R3, R5
I
II
III
N51
I47
Q50
5'
3'
D
R7
I
II
III
N51
R54
S50
5'
3'

water molecules. Arg31 and Arg53, however, contact the DNA backbone in slightly different positions in the three complexes (Gehring *et al*, 1994 A).

1.4 DNA binding specificity of homeodomains

The structural studies described have revealed a common mode of folding in solution and of DNA binding by homeoproteins. The amino acids identified as being critical to these processes have been observed to be highly conserved among homeodomains of all classes. The extremely variable biological roles of this large group of proteins, many of which have been shown to act as transcription factors (McGinnis and Krumlauf, 1992), had led to the expectation that biological specificity would be dictated by DNA binding specificity. However, the DNA binding properties of some homeoproteins have been analysed by DNase I footprinting, mobility shift assays and transactivation assays (Laughon, 1991, Gehring *et al*, 1994 B and Kalionis and O'Farrell, 1993). The majority of the homeodomains characterised recognise DNA sequences containing a 5'-TAAT-3' core motif (Gehring *et al*, 1994 A). These include members of the *Drosophila* Antennapedia (BS2 $C_{-3}T_{-2}C_{-1}T_0T_1T_2T_3T_4C_5G_6G_7T_8A_9A_{10}T_{11}C_{12}T_{13}C_{14}$, Muller *et al*, 1988) and Bithorax gene families as well as a number of more divergent homeodomains including Engrailed, Even-skipped, Paired and Bicoid (5'-CCCTAATCT-3' Driever and Nusslein-Volhard, 1989, Gehring *et al*, 1994 A). Only Rat Thyroid Transcription Factor 1 (TTF-1) a member of the divergent NK2 class of homeoproteins shows a peculiar DNA binding specificity, preferentially recognising sequences containing a 5'-CAAG-3' core motif (Oligonucleotide C, $C_{-2}A_{-1}G_0T_1C_2A_3A_4G_5T_6G_7T_8T_9$, Guazzi *et al*, 1990 and Damante *et al*, 1994).

Genetic studies were necessary to expand on the information derived from the structural data in order to reveal the mechanisms by which these highly conserved proteins may distinguish their specific target genes from those of other homeoproteins.

1.4.1 N-terminal arm

The structural studies had revealed that the N-terminal arm grasps the DNA in the minor groove, amino acids in this region were therefore strong candidates for determining DNA binding specificity of homeoproteins (Gehring *et al*, 1994 A). Studies of chimaeric HOM/Hox proteins have demonstrated that the sequence of the N-terminal arm is indeed critical for *in vivo* specificity. This was clearly illustrated by chimeras generated between Antp and Scr. These two homeodomains differ in only five of the 60 residues and four of these differences are within the N-terminal arm. Significantly, when these four residues are derived from Scr the protein behaves like Scr *in vivo*, even when the remainder of the protein is derived from Antp. Importantly, the criteria used to distinguish Scr from Antp included several independent assays: transformations of the first instar larva and adult antenna, as well as the regulation of specific downstream genes (Mann, 1995).

Several TTF-1HD mutants were also generated in which amino acids in the N-terminal arm were changed to those of Antp. A striking change in binding properties was observed with the mutant TTF-1HD(QTY) in which val6, Leu7 and Phe8 were changed to Gln, Thr and Tyr respectively (as Antp). TTF-1HD(QTY) prefers a 5'-TAAG-3' containing sequence, even though it still binds efficiently to 5'-CAAG-3' containing sequences. The corresponding amino acids of TTF-1HD were introduced into the wild-type Antp protein to produce AntpHD(VLF). As predicted, this mutant binds preferentially to 5'-CAAT-3' containing sequences (Damante *et al*, 1996).

The trajectory of the N-terminal arm of Eve differs from that of the En and Antp structures, a difference that may be due to tyrosine 4 and its interaction with the adenine of base pair 3 ($T_1A_2A_3T_4$). Tyr4 is unique to the Eve class of homeodomains (En contains a Proline and Antp a Glycine). Due to variability in the N-terminal paths, the highly conserved residue Arg3 interacts with DNA differently in the three structures. The identity of residue 4 may therefore be particularly important in determining the path of the N-terminal arm and the consequent DNA binding preferences (Hirsch and Agaarwaal, 1994).

1.4.2 Recognition helix

AA50

Numerous contacts had been observed between the recognition helix and both strands of the DNA in the structural studies. Treisman *et al*, mutated amino acid 50 of the Paired homeodomain and analysed the DNA binding specificities of the wild-type and mutant proteins (Treisman *et al*, 1989). The unmodified Prd protein does not recognise sequences specifically bound by the homeodomain protein Ftz, nor those specifically bound by Bicoid. When a Glutamine occupies this position, as in Ftz, the mutated Prd is able to bind to the Ftz sites (5'-TAATGG-3'). When a lysine occupies this position, as in Bcd the Prd mutant recognises the Bcd sites present in the hunchback promoter (5'-TAATCC-3') (Treisman *et al*, 1989). In addition, a Gln to Lys mutation at position 50 in the Ftz homeodomain causes a change in the recognition sequence from 5'-TAATGG-3' to 5'-TAATCC-3' (Damante *et al*, 1994). Base pairs contacted by residue 50 of Prd, Ftz and Bicoid are therefore located to the 3' side of the core motif 5'-TAAT-3'.

Similarly, a mutated TTF-1HD, in which residue 50 has been changed from Gln to Lys, binds efficiently to a mutated oligonucleotide C where the sequence 5'-CAAGTG-3' has been changed to 5'-CAAGCC-3'. These data indicate that the residue in position 50 of the TTF-1HD also makes contacts within the dinucleotide 3' to the motif 5'-CAAG-3'. This finding definitively demonstrates that TTF-1 HD, although possessing a peculiar DNA-binding specificity, recognises its cognate sequence using a geometry of interaction similar to that of other homeodomains (Damante *et al*, 1994).

Surprisingly, further mutant studies revealed that the use of Gln50 of TTF-1HD to recognize the base pair 3' downstream of the core is subordinate to the presence of the core motif 5'-CAAG-3' as binding of TTF-1HDAA50 mutants are not altered in the context of a 5'-TAAT-3' core. This suggests that in the absence of the 5'-CAAG-3' motif the amino acid at position 50 cannot be properly positioned in order to play a role in sequence-specific recognition (Damante *et al*,1994).

AA54

Mutations in oligonucleotide C that change the 5'-TCAAGTG-3' motif to 5'-TTAAGTG-3' (Cant4), 5'-TCAATTG-3' (Cant2) or 5'-TTAATTG-3' (Cant1) show an affinity for TTF-1HD lower than that of wild-type oligonucleotide C, confirming the relevance of the bases on either side of the central AA. TTF-1HD seems to be particularly sensitive to the base at position 5 ($T_1C_2A_3A_4G_5T_6G_7$) since mutation in this base (Cant2) causes the most significant decrease in binding activity. Antp binds with greatest affinity to Cant1 demonstrating that the bases flanking the conserved AA dinucleotide are those that allow the Antp to discriminate between oligonucleotide C and BS2. Mutants of the BS2 oligonucleotide confirm the importance of these 2 positions. In the BS2 context position 5 is still important for TTF-1 binding as a BS2 mutant in which only the T at position 5 has been changed for a G, is recognised much better by TTF-HD than wild-type BS2. Thus position 5 plays a major role in dictating binding for either TTF1-HD or Antp HD.

There were several factors to implicate amino acid 54 in the specific selection of a G in position 5 of the core motif of the TTF-1 binding. A tyrosine at this position is unique and conserved in all members of the NK class of homeodomains whereas Antp class homeodomains have a conserved methionine at this position. Structural studies of Antp/DNA and Matα2/DNA complexes demonstrate that residue 54 contacts DNA and a modelling study shows that Tyr54 of TTF1-HD is in an excellent position to contact DNA in the major groove very close to the 5'-CAAG-3' motif (Damante *et al*, 1994).

In order to test the relevance of Tyr54 for the recognition of 5'-CAAG-3' by TTF-1HD, the mutant TTF-1HD(Met54) was constructed. The DNA binding specificity of wild-type and mutant homeodomains was evaluated using wild-type and mutant oligonucleotide C and BS2 sites. The data shows that independent of whether the flanking sequence is that of oligonucleotide C or BS2 the wild-type TTF-1HD prefers G at position 5 whereas TTF-1HD(Met54) prefers oligonucleotides bearing T in place of G at position 5. These findings demonstrate that the nature of the residue at position 54 of the TTF-1HD

determines a preference for specific nucleotides at position 5 of the DNA binding sites and hence contributes to the overall DNA binding specificity of this homeodomain (Damante *et al*, 1996).

To verify whether the nature of residue 54 is a determinant for the DNA binding specificity of other homeodomains the binding of Antp and En homeodomains, which bear Met and Ala respectively at 54, to mutant and wild-type oligonucleotide C and BS2 sites was studied. The presence of either Ala or Met in pos54 does not appear to affect the DNA binding specificity as no difference in DNA sequence selection between Antp and En homeodomains was detected. However, when Antp and En were mutated to Antp(Tyr 54) and En(Tyr54) both mutant homeodomains switched their preference towards sequences containing the 5'-TAAG-3' core motif. Position 54 of homeodomains is therefore of relevance in determining the DNA sequence recognises. In addition to tyrosine, methionine and alanine several other residues can be found in this position in different homeodomains, it is possible that such differences might contribute to dictating the *in vivo* target genes for the corresponding homeodomains (Damante *et al*, 1996).

Interaction of AA50 and AA54

Amino acids 50 and 54 lie very close to each other in the recognition helix and are part of a fluctuating, short-lived bonding network. These structural features suggest that interactions may occur between these two residues, so that only particular combinations may ensure correct binding properties to homeodomains. In fact about 400 homeodomains from different species have been sequenced thus far and a bias in the combinations of amino acids 50-54 could be observed. Most homeodomains contain Gln at position 50 but few contain Lys at the same position. Among the Gln50 containing homeodomains, more than 50% contain Met at position 54. However, none of the Lys50 containing homeodomains contain Met at position 54 (Pellizzari *et al*, 1997).

Antp HD contains Gln and Met at positions 50 and 54 respectively and at the same positions the Gsc HD contains Lys and Ala. When in the context of Antp or Goosecoid

homeodomains Lys50 is paired to Tyr54 or Ala54 and Gln50 is paired to Met54 the resulting proteins are able to discriminate efficiently among different DNA sequences. In contrast, in the presence of the pair Lys50 and Met54 both homeodomains show a reduced capability to discriminate among different DNA sequences. Sequence selection experiments performed in the context of the Goosecoid homeodomain also suggest that the presence of Met54 precludes the base-discriminating function of Lys-50. These experiments demonstrate a functional interference between base-contacting amino acids at positions 50 and 54 of homeodomains which may explain why the pair Lys50-Met54 is never found in natural homeodomains (Pellizzari *et al*, 1997). The TTF-1HD QTY-M54 prefers sequences that contain 5'-TAAT-3' and the Antp VLF-Y54 binds most efficiently to sequences containing 5'-CAAG-3'. There are however quantitative changes among the different mutants that cannot be explained entirely by the mutations introduced suggesting that there must be additional contributions to DNA binding preferences (Damante *et al*, 1996).

Structural studies have therefore revealed that a core of highly conserved amino acids leads to a common mode of folding of all homeoproteins, even when sequence identity is as little as 27%. These structural studies implicated the N-terminal arm and Helix III in the determination of DNA binding specificity and this has been confirmed by genetic analysis with AA50 and AA54 playing a key role.

1.5 The NK class of homeodomain proteins

In 1989 Kim and Nirenberg isolated four *Drosophila melanogaster* homeobox genes that are divergent from the Hox like genes by screening a genomic DNA library with oligodeoxynucleotides encoding part of the presumed recognition domain of homeobox proteins (Kim and Nirenberg, 1989). The amino acid sequences of NK-2, NK-3, and NK-4 homeoboxes are more closely related to each other (59-66% homology) than they are to other *Drosophila* homeoboxes (28-54% homology). The homeobox of *NK-1*, however is most closely related, in order of decreasing homology, to muscle segment homeobox, *zerknullt-1, NK-3,* and *distal-less* homeoboxes. Three of the genes, *NK-1, NK-3*, and *NK-*

4, comprise a cluster of homeobox genes located in the 93E1-5 region in the right arm of the third chromosome, whereas the fourth homeobox gene, *NK-2*, is located in the 1C1-5 region of the X chromosome (Kim and Nirenberg 1989). The expression patterns of these genes have been studied and functions identified by examining mutant phenotypes.

1.5.1 *NK-1*

This gene, also known as *S59*, has been shown by *in-situ* hybridisation to be expressed in a small number of segmentally repeated mesodermal cells approximately 2 hours postgastrulation (Dohrmann *et al,* 1990). Gradually, four groups of *S59* expressing mesodermal cells appear in each abdominal hemi-segment, each one giving rise to a particular somatic muscle after fusion with surrounding myoblasts. In addition to the mesoderm, *S59* is expressed in a subset of neuronal cells of the CNS and their precursors, as well as in cells of a small region of the midgut. Thus *S59* may play a role in specifying the identity of particular somatic muscles and neurons of the CNS (Dohrmann *et al,* 1990).

1.5.2 *NK-2*

Expression pattern of *NK-2*

Northern analysis shows that *NK-2* mRNA is present at the highest concentration in 3-6 hour *Drosophila* embryos and then progressively decreases during further embryonic development. *In situ* hybridisation has revealed that *NK-2* gene expression is initiated during stage 4 in a bilaterally symmetrical pair of longitudinal stripes, in the ventral half of the ventrolateral neurogenic anlage. *NK-2* is also expressed in regions of the midgut and hindgut progenitors. Following mesoderm invagination the stripes of *NK-2* expressing neurectodermal cells fuse and give rise to the medial and paramedial neuroblasts which continue to express *NK-2*. (Nirenberg *et al,* 1995)

Binding studies to oligodeoxynucleotides have demonstrated that the consensus nucleotide sequence for NK-2 homeodomain binding is TNAAGTGG (Nirenberg *et al,* 1995). These NK-2 homeodomain binding sites were identified in 2.2Kb of the upstream

region of the *NK-2* gene, which suggests that NK-2 protein, or a related NK class protein may be required to maintain *NK-2* gene expression. (Nirenberg *et al*, 1995)

Mutant studies of *NK-2*

Mutation of the ventral nervous system condensation defective (*vnd*) locus had been observed to cause a phenotype in which a sub-population of neuroblasts are lost at early stages of development and later segregation of neuroblasts is interrupted. The expression pattern determined for *NK-2* was shown to correlate with the observed phenotype of *vnd* mutants suggesting that *NK-2* may be the transcript produced by the vnd locus (Jimenez *et al*, 1995). Jimenez *et al*, suggest that *NK-2* may be one component of a genetic system which establishes co-ordinates for the specific activation of the proneural clusters. *NK-2* would be induced to translate and activate proneural genes, by positional cues laid down at earlier stages by dorsal-ventral patterning genes (Jimenez *et al*, 1995).

Structure of the homeodomain of NK2

Most members of the NK family of homeodomain proteins including NK-2, share the unusual characteristic of having tyrosine as the 54^{th} and serine or threonine as the 56^{th} amino acid residues of the homeodomain. NK-2 also has the rare feature of a histidine at position 52. The three dimensional structure of the NK2 homeodomain has, however, been shown by NMR to contain 3 α-helices, consisting of helix I from residues 10 to 22, helix II from residues 23 to 38, a turn of residues 39 to 41 and helix III, from residues 42 to 52. Helix II and III form a helix-turn-helix motif, thereby forming an overall structure analogous to that found in the previously studied homeodomains. Despite these parallels in the structures of various homeodomains, individual sequential variations do result in significant local structural differences (Tsao *et al*, 1995). Residues 54-60 of NK-2 show a high degree of disorder (as also observed in Ftz), long-range constraints for NK-2, however, were found between the side-chain methyl group Thr56 and the aromatic ring of Phe 20. These constraints help to place the OH group of the Thr56 in a position where it can form a hydrogen bond with the backbone carbonyl group of His52. These

observations suggest that residues 53 to 56 adopt a flexible helix structure whereas residues 57 to 60 are unstructured (Tsao *et al*, 1995).

1.5.3 *NK-4* (*tinman*)

Expression pattern of *tinman*

The *tinman* (*tin*) gene is first expressed at the blastoderm stage in all mesodermal primordia of the trunk region and this expression occurs immediately after the onset of expression of the mesoderm determinant *twist,* a helix-loop -helix protein which binds to E-box DNA motifs (Azpiazu and Frasch, 1993). Two distinct clusters of E-box regulatory sequences (E1 contains the CANNTG motif and E2 TCAAGTG), present upstream of the *tinman* gene are required for *tinman* activation in cells of the visceral mesoderm demonstrating that *tinman* is a direct transcriptional target of *twist* (Mi Lee *et al*, 1997). *Tinman* expression persists during gastrulation as the mesoderm invaginates ventrally. After a portion of the mesoderm has migrated dorsally beneath the ectoderm, two layers form. The inner layer of mesoderm contributes to the visceral muscles, while the outer layer gives rise to the somatic muscles. At this point *tinman* expression is diminished in ventral portions of the mesoderm but is maintained in the visceral mesoderm. It is also maintained in the heart precursors, which are located most dorsally between the somatic and visceral mesoderm. Shortly thereafter, *tinman* RNA is further localised to the pericardial and cardial precursor populations and this expression persists throughout embryogenesis (Azpiazu and Frasch, 1993).

Frasch (1995) observed that mesodermal cells which do not contact *dpp* (*decapentaplegic*) expressing ectodermal cells completely lose *tinman* mRNA, whereas the cells below *dpp*-expressing ectodermal cell maintain high levels. The temporal and spatial correlation of the *dpp* and *tinman* expression patterns and the fact that *dpp* encodes a signalling molecule suggested that *dpp* might be involved in *tinman* activation in the dorsal mesoderm. In embryos lacking *dpp* function *tinman* expression is not maintained in the dorsal mesoderm, all *tinman* expression disappearing after mesoderm migration. Conversely, in embryos where *dpp* expression was expanded towards the ventral midline

tinman expression also exhibited a ventral expansion in its domain of expression. Together these results show that dorsal ectodermal cells that express *dpp* are required to maintain *tinman* expression in the underlying mesoderm (Frasch, 1995)

Mutant studies of *tinman*

In wild-type embryos at stage 16-17 when the heart is morphologically differentiated, Zfh-1 (Zinc-Finger Homeodomain protein 1) is localised to all four rows of heart cells as well as the lymph glands (Bodmer, 1993). In order to determine the function of *tinman* in heart development, *tinman* mutants were incubated with an antibody against Zfh-1. In mutant embryos null for the *tinman* gene no Zfh-1 immunoreactivity can be detected in the region where the heart normally forms, indicating that *tinman* mutants have no heart. Inspection of younger embryos (stage 12/13) marked with anti-Zfh-1 antibodies indicate that not only is there a lack of heart differentiation but that the precursor cells of the heart, the cardioblasts, are also absent. This data is consistent with *tinman* playing a role in specification of cardiac tissue (Bodmer, 1993).

In order to determine a possible role of *tinman* in visceral muscle formation, PlacZ enhancer-trap lines with reporter gene expression in the early visceral mesoderm were expressed in a *tin* mutant background. Compared to the wild-type, *tinman* null mutants exhibited no reporter gene expression in visceral mesoderm or differentiated visceral muscles of the trunk region indicating that no visceral mesoderm was formed, however, visceral muscles around the hindgut form normally. The *tinman* gene, therefore, appears to be involved in the specification of the visceral muscles associated with the midgut but not the hindgut (Bodmer, 1993).

Potential targets of *tinman*

The experiments described suggest that following the dorsal restriction of *tinman* expression, the *tinman* gene initiates regulatory events leading to visceral mesoderm and heart formation (Frasch and Aziapu, 1993). One candidate for a *tinman* target gene is

bagpipe (bap), as *bagpipe* mRNA appears in dorsal mesoderm shortly after the restriction of *tinman* to these cells and *bagpipe* expression is dramatically reduced in *tinman* and *dpp* null mutants. Although *tinman* is initially expressed throughout the mesoderm primordium, it is only after gastrulation and restriction of *tinman* expression to the dorsal mesoderm by *dpp* that *bagpipe* becomes activated. The requirement for both *tinman* and *dpp* for activation of *bagpipe* has been demonstrated by experiments in which ectopic expression of *dpp* in early mesoderm caused the activation of *bagpipe* in mesoderm prior to gastrulation where *dpp* and *tinman* are co-expressed (Staehling-Hampton *et al,* 1994). Thus the regional restriction of *bagpipe* expression to the dorsal mesoderm is due in part to the combined actions of *tinman* and *dpp*.

D-mef2 is a MADS box gene required for the development and differentiation of *Drosophila* heart. Gajewski *et al*, 1997, have shown that D-mef2 expression in the developing *Drosophila* heart requires a novel upstream enhancer containing two Tinman binding sites, both of which are essential for enhancer function in cardiac muscle cells. Ectopic *tinman* expression activates the enhancer outside the cardiac lineage (Gajewski *et al*, 1997). The mutation of a GATA sequence in the enhancer changes its specificity from cardial to pericardial cells. Also, the addition of flanking sequences to the heart enhancer results in expression in the founder cells of a subset of body wall muscles. As *tinman* function is required for D-mef2 expression in both the cardial and the founder cells, these results define a shared regulatory DNA that functions in distinct lineages due to the combinatorial activity of Tinman and other factors that work through adjacent sequences (Gajewski *et al*, 1998).

1.5.4 *NK-3* (*bagpipe*)

Expression pattern of *bagpipe*

In situ hybridisation has shown that *bagpipe* starts to be expressed after the restriction of *tinman* mRNA to the dorsal mesoderm and that *bagpipe* expression is limited to the dorsal most cells of the mesoderm from its outset. In contrast to *tinman* expression in the dorsal mesoderm, which at stage 10 is continuous along the anterior posterior axis,

bagpipe expression is segmentally interrupted (Azpiazu and Frasch, 1993). The 11 patches of *bagpipe* expression are located in the region corresponding to segments 2-12 of the epidermis. A subset of cells from these clusters segregates to form visceral mesoderm that differentiates into gut musculature. In addition to the middle body region, *bagpipe* is expressed in the mesodermal cells at the proctodeum and stomodeum. These cells will develop into the visceral mesoderm of the hindgut and foregut, and continue *bagpipe* expression until late embryogenesis. Following stage 12, *bagpipe* expression appears in a subset of heart progenitors (Azpiazu and Frasch, 1993).

Mutant studies of *bagpipe*

In embryos mutant for *bagpipe*, visceral mesoderm formation is strongly disrupted. Most cells of the visceral mesoderm fail to differentiate properly and a portion of them are transformed into body wall musculature and gonadal mesoderm. Despite the expression of *bagpipe* in heart progenitors and cells which will contribute to the hindgut the development of both heart and hindgut are unaffected. In *tinman* mutant embryos, *bagpipe* is not activated in the dorsal mesoderm. Probably as a consequence, neither visceral mesoderm nor midgut musculature are formed in these mutants, and the absence of visceral mesoderm results in strong disruption of endoderm migration and midgut morphogenesis (Azpiazu and Frasch, 1993).

Control of *bagpipe* expression

The mesoderm of stage 10-11 embryos is organised into parasegmental repeats that are in exact register with the ectodermal parasegments. Each parasegment is subdivided into two domains along the anteroposterior axis. The mesodermal domains below the anterior compartments are termed 'A domains' and those beneath the posterior compartment 'P (Azpiazu et al. 1996).
domains'. The P domains express *bagpipe* and *serpent* (expression of this gene marks anlagen of the fat body) and give rise to visceral mesoderm derivatives that include midgut musculature and the fat body. The 'A domains' include the primordia of the cardiac mesoderm and the bulk of the somatic mesoderm and give rise to the heart and most of the body wall muscles. Azpiazu *et al*, have shown that *hedgehog (hh)* and

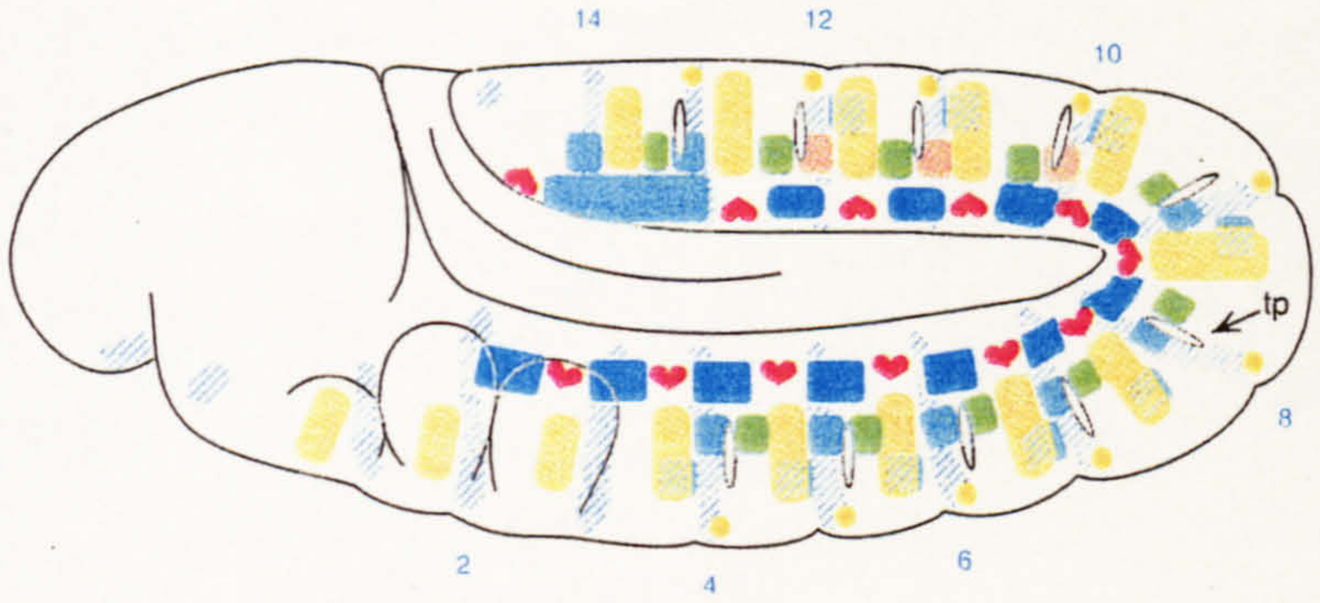
Figure. 6 A fate map of the trunk mesoderm of *Drosophila*

A. Schematic fate map of the mesodermal cell layer, also showing the prospective enctodermal parasegment boundaries, tracheal pits (tp) and the outlines of the future head segments. *En* stripes at the anterior border of each even PS are numbered. The postulated map is based partly on early gene expression patterns (e.g. *bap* for the midgut visceral mesoderm) and partly extrapolated back from later gene expression patterns (*eve* for the heart, high levels of *twist* for the somatic muscles, *slp* for the fat body) taking into account the cell and tissue movements observed during mesoderm development. The SGP are placed ventral to the visceral mesoderm, as judged by their location relative to *bap* expression. The fat body primordia are in the same relative dorsoventral position because they appear to be in serial homology to the SGP and because they lie outside the domain of *dpp* activity that directs visceral mesoderm development.

B. Fate maps of three representative segments including genes that regulate the subdivision of the mesoderm. The posterior border of the secondary (2°) dorsolateral fat body cluster is shown as a gradient because it is not known whether this primordium is completely included in the *eve* domain or extends into the *srp* domain. The ventral secondary fat body cluster is shown as overlapping with part of the somatic mesoderm. It is not known whether fat body and somatic mesodermal cells in this region are initially intermingled and later sort out, or whether they arise from distinct regions of the *wg* domain.

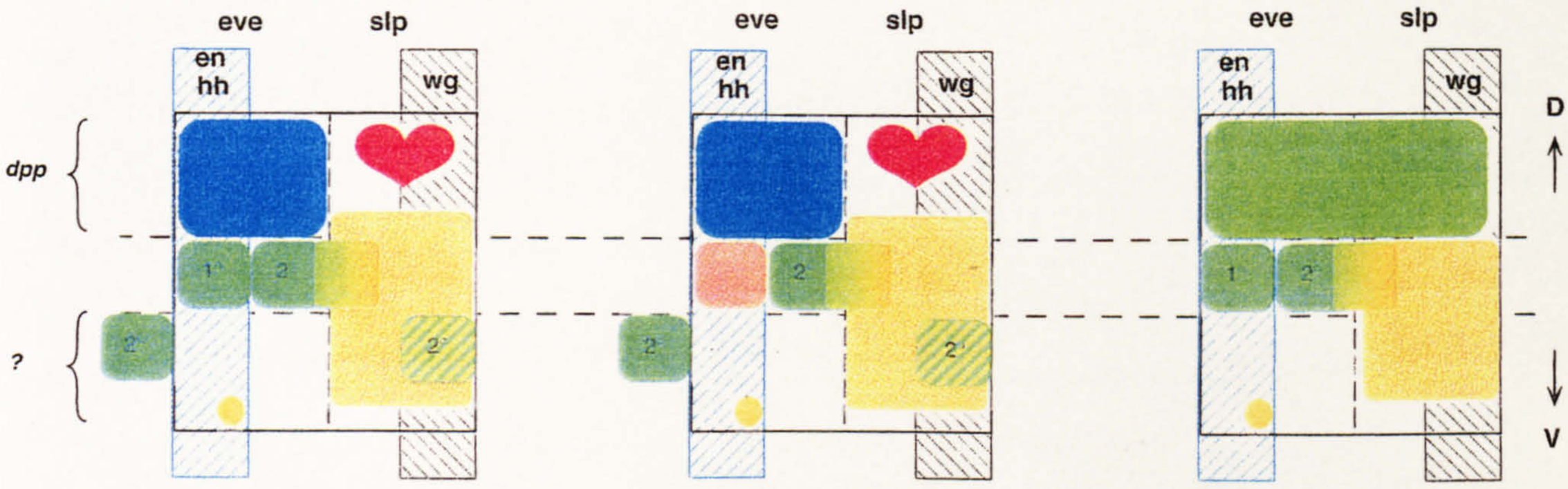
(Figure reproduced from Reichmann *et al*, 1998)

A


14
12
10
8
6
4
2
tp


somatic muscles
gonads
fat body
visceral muscles
heart
mesodermal glia
engrailed

B


eve
slp
en
hh
wg
dpp
?
1
2
2'
D
V
PS 4 - 9
PS 10 - 12
PS 13

wingless (wg) mediate pair-rule gene functions in the mesoderm, probably partly by acting within the mesoderm and partly by inductive signalling from the ectoderm. *hedgehog* is required for normal activation of *bagpipe* and *serpent* in anterior portions of each parasegment, whereas *wingless* is required to suppress *bagpipe* and *serpent* expression in posterior portions (Fig. 6) (Azpiazu *et al*, 1996).

Potential targets of *bagpipe*

In order to identify novel downstream target genes of *bagpipe* Lo *et al*, used an enhancer trap line which had demonstrated visceral mesoderm expression, to isolate a novel gene, *vimar* (visceral mesodermal armadillo-repeats) (Lo *et al*, 1998). A database search with the vimar protein sequence revealed sequence similarities to the human and bovine Smg GDS proteins. The Smg GDS protein is composed almost entirely of 11 tandem copies of a degenerate repeat sequence the Armadillo repeat, initially identified as a 42 amino acid repeat in the Armadillo protein, a *Drosophila* segment polarity gene product homologous to mammalian B-catenin. An alignment of vimar against the Arm repeat gene consensus sequence reveals that it is almost entirely comprised of 15 tandem copies of the Arm repeat, of which two are incomplete (Lo *et al*, 1998).

Expression of the vimar gene is first weakly detected in the mesoderm of stage 9 embryos with expression increasing considerably by stage 10 and becoming mostly restricted to the prospective hindgut and midgut visceral mesoderm, in which it is visible as 11 patches in the dorsal mesoderm. Expression can also been seen in the PNS precursor cells at this stage. Following germband retraction and dorsal closure the 11 patches have fused together and an increasing gradient of expression has formed from anterior to posterior with strong expression being maintained in the hindgut visceral mesoderm. At this stage expression is also detected in the CNS. Midgut visceral mesoderm expression starts to decrease so that by stage 17 it is barely detectable there, while it still remains strong in the hindgut visceral mesoderm. Expression is also detected in the CNS, PNS and the gonadal mesoderm at this stage (Lo *et al*, 1998).

The domains of vimar expression in the midgut and hindgut visceral mesoderm at stages 10 and 11 are very similar to those of *bagpipe*. Embryos homozygous for a *bagpipe* deletion show no detectable expression of *vimar* in the prospective midgut visceral mesoderm cells, while expression in the prospective hindgut remains intact. Thus, *bagpipe* is required for the expression of the *vimar* gene in the prospective midgut visceral mesoderm cells but not in the prospective hindgut visceral mesoderm (Lo *et al*, 1998). This is consistent with the observation that mutation of *bagpipe* has no detectable effect on the development of hindgut mesoderm (Aziapu and Frasch, 1993). These studies place *vimar* downstream of *bagpipe* but further studies will be necessary to determine whether *vimar* is a direct target of *bagpipe*.

A study of gonad development has revealed a second potential target of *bagpipe*. *Drosophila clift* is a gene required for the maintenance of somatic gonadal precursor (SGP) cell fate. SGP cells originate in regions just ventral to the visceral mesoderm precursors, near the boundary of *dpp* expression in overlying ectoderm. These SGP cells, like other dorsal mesodermal cells require *tinman* function (Boyle *et al*, 1997). In *bagpipe* null mutants the domain of *clift* expression is expanded; whereas the total number of SGP cells averages 25-37 in a 7 hour AEL (after egg laying) wild-type embryo, in *bagpipe* mutants there are 40-55 cells. Although it is difficult to determine precisely the position of the extra SGP cells they often appear in slightly more dorsal positions in the mesoderm compared to wild-type (Boyle *et al*, 1997). This would be consistent with extra SGP cells being recruited from cells that would normally become visceral mesoderm, which lies just dorsal to the SGP cells. Thus, the dorsoventral position of SGP specification is refined by negative interaction with the dorsally adjacent visceral mesoderm precursors and these results suggest that it may be wild-type *bagpipe* function which causes the repression SGP cell development although this effect is not necessarily direct (Boyle *et al*, 1997).

1.6 Vertebrate members of the NK class of homeobox genes

The tissue specific expression patterns and mutation analyses of the NK class genes in *Drosophila* demonstrated that this class of homeobox genes has a role in specification of

tissue type; the identification of vertebrate relatives was therefore a priority. Figure 7 illustrates an alignment of the three homology domains of the many of the vertebrate NK class homeobox genes.

1.6.1 Sequence comparison

The greatest homology between all the genes is obviously within the homeodomain. The tyrosine at position 54, a position which influences DNA binding site to position 4 of a $T_1A_2A_3T_4$ motif in other homeodomain proteins, is unique among the *Drosophila* NK-2, NK-3 and NK-4 proteins and their close relatives (Damante et 1996, and Pellizari *et al*, 1997, Lints *et al,* 1993). Along with the homeobox there are two additional regions of homology; the first is located carboxy-terminal to the homeodomain and consists of a 17 amino acid motif that is found in NK2 and its relatives with a divergent form appearing in NK3 sequences. This domain contains a central cluster of invariant hydrophobic amino acids (VPVLV). The second region of homology is a decapeptide located near the amino-terminal of some members of the class, including *tinman* and its relatives in vertebrates.

1.6.2 Nomenclature

Each vertebrate NK class homeobox gene has been assigned a name by virtue of its sequence homology in the conserved domains to one of the four *Drosophila* genes, followed by an additional number relating to their order of discovery. Expression and mutation data, however, have revealed that there is not a strict correlation between conservation of sequence and of function. One example is murine Nkx-2.5, which at the amino acid level is most related to NK-2, but its heart specific expression and disrupted heart formation in mouse knockouts suggest a function in heart formation, a function more related to that of NK-4 than NK-2 (Lyons *et al*, 1995). It appears that in vertebrates there is a large family of NK like homeobox genes, not four separate families that correspond with the four genes observed in *Drosophila*. Thus the nomenclature of the vertebrate NK genes is somewhat academic and does not give a clear picture of the relationships between *Drosophila* and vertebrate genes.

Figure. 7 Comparison of the Nkx-2.5 'TN domain, homeodomain and 'NK2' domain sequences with a range of NK class genes.

Gene names are prefixed with a species abbreviation: m, mouse; Dro, *Drosophila*; p, *planaria*; ch, chick; X, Xenopus; h, human; D.R. zebrafish danio rario; r, rat. All spequences are compared to mNkx-2.5 and identities are indicated by a dot. Over-lining at the top of the homeodomain section indicates predicted homeodomain helices.

Sequence references; Murine Nkx-2.5, Lints *et al*, 1993, Human csx, Turbay *et al*, 1996, chick Nkx-2.5, Schulthesis *et al*, 1996, *Xenopus* Nkx-2.5, Tonissen *et al*, 1994, d.rario nkx-2.3, nkx-2.5 and nkx-2.7, Lee *et al* 1996, *Xenopus* Nkx-2.3, Evans *et al*, 1995, chick Nkx-2.8, Brand *et al*, 1997, murine Nkx-2.6, Lints *et al*, 1993, NK-1, NK-2, NK-3 and NK-4, Kim *et al*, 1989, murine Nkx-2.2, Hartigan *et al*, 1996, *Xenopus* Nkx-2.2, XXX, 1996, Murine *Bapx1*, Triboli *et al*, 1997, murine Nkx-3.1, Bieberich *et al*, 1996, human Nkx-3.1, Prescott *et al* 1998, human bapx1, Yoshiura *et al*, 1997, *planaria bagpipe*, Balavoine *et al*, 1996, *planaria* Nkx-3.2 and Nkx-3.3, Nicolas *et al*, 1999, murine Sax1, Schubert *et al*, 1995, murine Nkx-5.1 and Nkx-5.2, Bober *et al*, 1994, lox 10, Nardeli-Haefiger *et al*, 1993, rat ttf-1, Guazzi *et al*, 1990, human ttf-1, Hamdon *et al*, 1998, *planaria* dth-1 and dth-2, Garcia-Ferandez *et al*, 1991, human Nkx-6.1, Inoue *et al*, 1997.

```
    mNKX-2.5  TPFSVKDILNL
  Ch.Nkx-2.5  ...........
    Xnkx-2.5  ...........
    XNkx-2.3  ...........
d.r Nkx-2.3   .........K.
d.r Nkx-2.7   .........K.
 Ch Nkx-2.8   .....E...S.
 Dro tinman   ..........M
Mus Nkx-3.1   .S.LIQ...RD
  Mus Bapx1   ....IQA...K
Dro bagpipe   ....IN...TR
 Pn Nkx-3.2   ....IQA...K
    Mus Sax1  IS...L...DP
     Hum Csx  ...........
    Hum ttf-1 .....S...SP

HOMEODOMAIN              helix I          helix II         helix III

    mNkx-2.5  RRKPRVLFSQAQVYELERRFKQQRYLSAPERDQLASVLKLTSTQVKIWFQNRRYKCKRQR
     Hum Csx  ...........................................................
  Ch.Nkx-2.5  ..............................H..N.........................
    XNkx-2.5  ........................K.....H..N.........................
d.r Nkx-2.5   ................Q.....K.....H..N.........................
    XNkx-2.3  ............F.................EH..NS.......................
d.r Nkx-2.3   ............F.................EH...T.......................
d.r.Nkx-2.7   R.......................K.....H..N.........................
 Ch Nkx-2.8   ..........T..L.........K....L..EH..N..Q.....................
              RKGGQ.R..ND.TI...KK.ET.K...P...KR..KM.Q.SER...T.....A.WR.LK
      NK-2    K..R....TK..T.......R..........EH...LIR..P.........H...T..AQ
    mNKx-2.2  K..R.....K..T.......R..........EH...LIR..P.........H...M..A.
    XNKx-2.2  K..R.....K..T.......R..........EH...LIR..P.........H...M..A.
  Dro tinman  K....V...Q...L...C..RLKK..TGA..EII.QK.N.SA.............S..GD
   Mus Bapx1  KKRS.AA..H...F......NH.....G...AD..AS....E.............T..RQ
Mus Nkx-3.1   QKRS.AA..HT..I....K.SH.K.......AH..KN....E.............T..KQ
    Hum Bap   KKRS.AA..H...F......NH.....G...AD..AS....E.............T..RQ
     Pn Bap   KKRT.AA..HT.........GH.....GS..AE..RS.R.SE..I..........T.KRQ
  Pn Nkx-3.2  KKRS.AA..H...F......NH.....G...AD..AS....E.............T..RQ
  Pn Nkx-3.3  KKRS.AA..H..........SL.....G...AA..AS....E.............T..KQ
Dro bagpipe   KKRS.AA..H...F......A......G...SEM.KS.R..E.............T..KQ
      NK-1    K..R....TK..T.......R..........EH...LIR..P.........H...T..AQ
    Mus Sax1  P.RA.TA.TYE.LVA..NK.RAT....VC..LN..LS.S..E...........T.W.K.N
 Mus Nkx-5.1  KK.T.TV..RS..FQ..ST.DMK....SS..AG..AS.H..E...........N.W...L
 Mus Nkx-5.2  KK.T.TV..RS...Q..ST.DMK....SS..AC...S.Q..E....T......N.W...L
      lox 10  ...R.I......I.......R..K.......EH..TFIG..P.........H...T.KSK
    Hum ttf-1 ...R.....................K.......EH...MIH..P.........H...M...A
     Pn Dth-1 K..R.....KK.IL....H.R.KK.......EH..NLIG.SP.........H...M..AH
     Pn Dth-2 ...R.I......I..........K.......EH..NLIN..P.........H......S.
 Hum Nkx-6.1  .KHT.PT..GQ.IFA..KT.E.TK..AG...AR..YS.GM.ES..........T.WRKKH


"NK2" DOMAIN
    mNkx-2.5  PPARRIAVPVLVRDG
     Hum Csx  ...............
  Ch Nkx-2.5  ..P............
    XNkx-2.5  ..P............
   XNkx-2.3.  ..P..V.........
d.r Nkx-2.3   ..P..V.........
d.r Nkx-2.7   ..P...S........
  Ch Nkx-2.8  L.P..V.........
   Mus Bapx1  .A.KKV..K.....D
  Pn Nkx-3.2  GA.KKV..K.....D
  Pn Nkx-3.3  S....V.IR.....D
Dro bagpipe   GASK.VPIQ....ED
   hum ttf-1  QSP..V......K..
```

1.6.3 Characteristics of the vertebrate *NK* homeobox genes

As illustrated by figure 7, the majority of the vertebrate *NK* class genes have been assigned to an *NK-2* family. A *Xenopus NK-1* has been cloned (*XNkx-1.1*) but has not yet been characterised (Bober *et al,* 1994). One of the best characterised vertebrate *NK* class genes is *Nkx-2.1*, which is also known as *ttf-1* (Thyroid specific Transcription factor) (Guazzi *et al,* 1990). Rat *ttf-1* is required for the specific expression of the thyroglobulin and thyroperoxidase gene promoters in differentiated thyroid cell lines, as shown by cell transfection assays. *In situ* hybridisation has revealed *ttf-1* expression in the thyroid anlage, the lung and in specific regions of the brain (Lazzaro *et al*, 1991). Several other *NK-2* related genes identified in mouse and *Xenopus* are expressed in neural tissue including m*Nkx-2.2*, X*Nkx-2.2* (Saha *et al*, 1993, Price *et al,* 1992). Two murine *NK* class homeobox genes have been classified as *Nkx 5.1* and *5.2* as they have sequence motifs similar to more than one of the *Drosophila* genes, their expression is restricted to specific regions of the brain and the otic vesicle (Bober *et al* , 1994, Rinkwitz-Brandt *et al*, 1995).

A second group of *NK-2* related genes show a distinct pattern of expression, their transcripts being detected principally in regions of the embryo associated with heart formation. There are now many genes in this group including m*Nkx-2.5*, m*Nkx-2.6,* X*Nkx-2.3*, X*Nkx-2.5*, ch*Nkx-2.5*, ch*Nkx-2.8*, D.r *nkx-2.3* and D.r *nkx-2.7*. The most extensively characterised gene of this group is murine *Nkx-2.5*.

1.6.4 Murine *Nkx-2.5*

Expression pattern of murine *Nkx 2.5*

Nkx 2.5 transcripts are first detected at early headfold stages in myocardiogenic progenitor cells of the mouse embryo (Lints *et al*, 1993). Expression precedes the onset of myogenic differentiation, and continues in cardiomyocytes of embryonic foetal and adult hearts. Transcripts are also detected in future pharyngeal endoderm, the tissue believed to produce a signal, which induces heart formation. Expression in endoderm is only found laterally, where it is in direct apposition to promyocardium. After foregut closure, *Nkx 2.5* expression is limited to the pharyngeal floor, dorsal to the developing heart tube. The

thyroid primordium, a derivative of the pharyngeal floor, continues to express *Nkx 2.5* after transcript levels diminish in the rest of the pharynx. *Nkx 2.5* transcripts are also detected in lingual muscle, spleen and stomach (Lints *et al*, 1993).

Overall, the patterns of expression of *XNkx-2.3* and *XNkx-2.5* are remarkably similar to *mNkx-2.5*, and all imply that these factors play a role in commitment to and/or differentiation of the myocardial lineage. The expression of this group in cardiogenic regions is reminiscent of the expression of *tinman* in cardiogenic mesoderm. The expression of m*Nkx-2.5* in the stomach mesoderm and of *XNkx-2.3, XNkx-2.5* and *mNkx2.5* in many derivatives of the visceral mesoderm also correlates well with the expression of *Drosophila tinman* in visceral mesoderm.

Mutation of *Nkx-2.5*

To investigate the functions of *Nkx-2.5* mutant mice completely null for *Nkx-2.5* were generated and analysed. The null mutation did not eliminate the heart cell lineage suggesting that the role of *Nkx-2.5* is different to that of *tinman* which specifies the precardiac mesoderm in *Drosophila* (Tanaka *et al*, 1999). In contrast to an earlier study of an insertion mutation of *Nkx-2.5* (Lyons *et al*, 1995) initial heart looping was also observed to occur normally. However, homozygous null mutants showed arrest of cardiac development after looping and there were severe defects in vascular formation and hematopoesis in the mutant yolk sac. These mutants also revealed that *Nkx-2.5* regulates the expression of ANF (atrial natriuretic factor) and BNP (brain natriuretic peptide), N-myc, MEF2C and eHAND, transcription factors, which are involved in heart development. The data presented by Tanaka's group therefore demonstrates that *Nkx-2.5* is essential for normal heart development (Tanaka *et al*, 1999).

If *Nkx-2.5* is a true functional equivalent of *tinman* it may seem surprising that there is any heart formation at all. However, two other genes closely related to *Nkx-2.5*, *Nkx-2.3* and *Nkx-2.6*, have been isolated in the mouse which are also expressed in heart. It may be

Fu et al recently conducted studies in which DNA encoding dominantly acting repressor derivatives of XNkx-2.3 and XNkx-2.5 (XNkx-2.3EnHD and XNkx-2.5EnHD) were injected into developing *Xenopus* embryos. Injection of either construct resulted in a significant occurrence of embryos mutant for staining of cardiac mesoderm markers cardiac troponin C and myosin light chain chain C. Interestingly, coinjection XNkx-2.3EnHD and XNkx-2.5EnHD constructs produced a synergistic effect, resulting in significantly higher frequency of mutant phenotypes than obtained with either construct injected alone. Rescue of mutant phenotypes was effected by coinjection of either wild-type XNkx-2.3 or XNkx-2.5. Vertebrate *tinman* homologs are therefore required for the earliest stages of *Xenopus* heart formation and act in a functionally redundant manner (Fu et al, 1998).

that, as observed in other gene families, there is partial functional redundancy between individual *NK2* family members (Lyons *et al* 1995).

Studies of the vertebrate *NK2* genes had revealed genes that appear to be related to *Drosophila tinman* in both sequence, expression and function. It was, therefore, possible that vertebrate relatives of the *Drosophila bagpipe* gene, which lies downstream of *tinman* and plays a role in the differentiation of visceral mesoderm, may also be identified. At the commencement of this thesis no vertebrate *NK3* relatives had been identified; in the four years since, *NK3* genes have been identified in toads, mice, humans and *planaria*. The focus of this thesis is the cloning and molecular characterisation of a *Xenopus* relative of *Drosophila bagpipe* and, therefore, each of these genes will be described.

1.6.5 Human *NKX-3.1*

He *et al*, isolated a human *NKX3.1* gene from a prostate tissue library using an EST which had demonstrated prostrate specific expression. The sequences of the NK3 class genes are illustrated in figure 7. Northern analyses using human tissues revealed high levels of expression in prostrate and at much lower levels in testis. *NKX3.1* mRNA was not detected in several other adult tissues including brain, kidney, small intestine, pancreas, heart, liver, lung, thymus, spleen, placenta, colon, lymphocytes and ovary. Studies of *NKX3.1* expression in a variety of human cell lines revealed that *NKX3.1* expression is seen in the hormone-responsive, androgen receptor-positive prostate cancer cell line, but not in androgen receptor-negative cell lines nor in eleven other cell lines of varied tissue origin (He *et al*, 1997). These observations were investigated further by comparing *NKX3.1* expression in cells grown in standard medium and in the absence of androgens. The removal of androgens from the medium resulted in reduction of *NKX3.1* expression to an undetectable level. These results led He *et al*, to suggest that *NKX3.1* is part of the prostrate cells primary response to androgen stimulation and that the gene is a candidate for playing a central role in the differentiation of normal prostatic tissue and potentially the reversal of normal differentiation seen during cancer progression (He *et al*,

1997). The *NKX3.1* gene has been mapped to chromosome band 8p21, a region frequently reported to undergo a loss of heterozygosity associated with tissue dedifferentiation and loss of androgen responsiveness during the progression of prostrate cancer (He *et al*, 1997).

1.6.6 Murine *Nkx-3.1*

Screening a day 17.5pc murine embryonic cDNA library with a probe produced using sequence information from the human cDNA isolated murine *Nkx-3.1*. The sequence of the Nkx-3.1 TN domain and homeodomain are featured in the alignment of figure 7, mNkx-3.1 does not appear to contain a recognisable NK2 box (Sciavolino *et al*, 1997).

RNase protection analysis revealed expression of *Nkx-3.1* during late stages of gestation in the male urogenital sinus, prostatic buds, and testis. In adult males, expression of Nkx-3.1 is increased during sexual maturation and is dramatically reduced upon castration, suggesting that androgens are required for maintenance of expression (Sciavolino *et al*, 1997). During embryogenesis expression of *Nkx-3.1* was detected by in situ hybridisation in the developing urogenital sinus at day 14.5 p.c, and in the ventral prostatic buds at day 17.5 p.c. At day 14.5 p.c. expression of *Nkx-3.1* was observed in the out-buddings of the pelvic region of the urogenital sinus, which are responsible for the formation of the prostate gland, with low levels of expression in the prospective urethra. Expression was also observed at both day's 14.5 p.c. and 17.5 p.c. in the developing testes. Although female sexual organs fail to express *Nkx-3.1* at a detectable level at any stage expression of Nkx-3.1 is not restricted to the male urogenital system; expression was also observed in the dorsal aorta and in the kidney (Sciavolino *et al*, 1997).

1.6.7 Murine *Bapx1*

A mouse genomic phage library was screened using a *Drosophila bagpipe* homeobox probe and a Pst fragment from the one positive phage was used to screen an E11.5 murine cDNA library. The novel homeobox gene isolated contained a TN domain, a homeobox

identical to *XBap*, and an NK2 domain and was designated *Bapx1* (Fig.7). *Bapx1* maps to the proximal end of chromosome 5 in mouse (Triboli *et al*, 1997).

Bapx1 transcript distribution during embryonic development was analysed by RNA *in situ* hybridisation in whole mount and sectioned embryos. *Bapx1* expression is first detected in the second or third most recently formed somite and in the lateral plate mesoderm overlying the future gut endoderm at E8.0-E8.5. During rotation of the embryo strong *Bapx1* expression is seen in the lateral plate mesoderm that folds to surround the epithelium of the gut and in the sclerotome portion of the somite. This expression follows a rostrocaudal gradient of intensity with the more rostral (developmentally older) regions showing greater intensity than the more caudally positioned structures. By E9.5, a novel domain of *Bapx1* expression is seen in the branchial arches, with greatest intensity seen in the inferior portion on the first branchial arch (Triboli *et al*, 1997).

Bapx1 transcript distribution at E10.5 is similar to earlier stages and this expression pattern persists into later stages. By E12.5, *Bapx1* expression is also detectable in the developing limbs where expression is restricted primarily to cartilaginous condensations. The earlier expression in the branchial arches is restricted by E12.5 to the precursor of Meckels cartilage. At this stage, *Bapx1* expression in the gut mesentary is localised to the ventral mesoderm of the stomach and to discreet patches surrounding the hindgut. In older embryos (E14.5-E16.5) *Bapx1* transcripts are detected qualitatively in all the same structures as earlier stages, but by E16.5, there is quantitatively a significant decrease in the level of expression. Within the axial skeleton, expression is strongest in the most caudal prevertebrae and within the developing limbs the strongest expression of *Bapx1* is observed in the digits (Triboli *et al*, 1997).

1.6.8 Human *BAPX1*

The human relative of mouse *Bapx1* was isolated using a fragment of the mouse gene to screen a human genomic library and a craniofacial cDNA library (Yoshira *et al*, 1997). The alignment of the sequence in figure 7 demonstrates that Human BAPX1 contains a

TN domain and an NK2 domain. Linkage mapping was used to localise *BAPX1* to 4p16, a region to which the gene for Ellis-van Creveld syndrome has also been mapped. This syndrome is an autosomal-recessive disorder with skeletal anomalies including short-limbed, disproportionate dwarfism, polydactyl, hypondia and congenital heart disease. Thus, based on position and the expression pattern of the mouse homologue *Bapx1* is a candidate gene for involvement in Ellis-van Creveld syndrome (Yoshira *et al*, 1997).

1.6.9 *Xenopus Bagpipe* (*XBap*)

Newman *et al*, isolated *Xenopus Bagpipe* from a screen of an adult heart cDNA library using the *Nkx-2.5* homeodomain as a probe. The XBap sequence contains a TN domain and a degenerate version of an NK2 box (Newman *et al*, 1997).

RNase protection analysis first detected embryonic expression of *XBap* at the late neurula stage of development. The level of *XBap* increases slowly until stage 30 when the transcript abundance climbs sharply. *XBap* RNA levels then remain approximately constant through the tailbud stages (stages 30-44). Whole mount and sectioned in-situ hybridisation studies were used to localise the expression of *XBap* within the embryo and revealed two main domains of expression, first in the developing face and later in the anterior gut tissues (Newman *et al*, 1997).

Craniofacial expression of *XBap* is first detected in the early tailbud embryo (stage 30) and becomes clearly defined slightly later as distinct, bilaterally symmetrical bars, on either side of the cement gland and flanking the future mouth opening. As development proceeds, this region contracts until *XBap* expression is concentrated in two regions immediately ventro-posterior to the stomodeum. At stage 32 staining is detected below the floor of the developing pharyngeal cavity. This domain is consistent with the region where the basihyobranchial cartilage will develop. Interestingly, unlike most other cartilage of the facial region, the anuran basihyobranchial cartilage is not derived from the cranial neural crest but has its origins in the anterior mesoderm. At stage 37 *XBap* expression is located in two patches ventro-lateral to the mouth cavity. By stage 42 *XBap*

staining marks the anterior and lateral regions of the mouth cavity as well as a narrow patch of cells below the posterior pharyngeal cavity, all of these regions lie within the mesenchymal layer (Newman *et al*, 1997).

In tailbud embryos, *XBap* transcripts are detected in bilateral regions dorsoposterior to the developing heart. Although expression is initially located high on the flank, the area of expression shifts ventrally and broadens at the base as development proceeds. Expression is also observed to diminish on the right side of the embryo until it finally disappears at about stage 40. Further examination by taking sections through stage 37 embryos show that *XBap* expression is localised to a thin layer of cells below the surface ectoderm. By stage 41 the staining has extended much deeper below the surface of the embryo, and a thin layer of *XBap* expressing cells is separated from the lumen of the gut by a thick layer of non-expressing endodermal tissue. The *XBap* expressing cells are located at the outer edges of the gut tube, which will go on to form the gut musculature (Newman *et al*, 1997).

In situ hybridisation studies failed to detect any *XBap* expression in the posterior gut region in contrast to the expression pattern of *bap* in the *Drosophila* embryo. This was confirmed by carrying out RNase protection assays on head, middle and tail segments of stage 34 embryos. These experiments confirmed that at this stage there is no expression of *XBap* in the posterior region where the hindgut is developing. RT-PCR assays carried out on foregut and hindgut sections of stage 46 tadpoles, however, reveal approximately equal levels of *XBap* expression in the two regions indicating that *XBap* expression extends into the hindgut tissues during the later stages of embryogenesis (Newman *et al*, 1997).

Expression of *XBap* in adult tissues was assayed using RT-PCR. Significant expression was detected in the kidney, pancreas, spleen and stomach with lower levels detected in intestine and skeletal muscle, very low levels in the heart and lung and no expression was detected in the liver (Newman *et al*, 1997).

The expression of *XBap* in the developing gut muscle is approximately equivalent to *bap* expression in the *Drosophila* embryo. *XBap* is not expressed in the heart anlage, however, despite it's expression in this region mutation of *bap* has no effect on development of the dorsal vessel in *Drosophila*. The expression of *XBap* in the developing musculature of the midgut suggests that one role of *Drosophila bagpipe* has been evolutionarily conserved, whereas, expression in the cartilage of the jaw appears to represent a new role for *XBap* (Newman *et al*, 1997).

1.7 Aims of the project

This project commenced in 1995 at which time there were no vertebrate homologues of *Drosophila bagpipe*. The identification of a *Xenopus* relative of *bagpipe*, which may potentially have a conserved role in visceral mesoderm patterning was therefore of great interest. The primary aim of this project was to isolate and study any *Xenopus* homologues of *bagpipe*.

During the course of this project a study of *XBap* was published, giving a detailed expression pattern for the *XBap* gene. For this reason, the remainder of the thesis studies have examined the DNA binding and transcriptional activity of the XBap protein.

Chapter 2

Materials and Methods

2.0 Chapter 2: Materials and Methods

2.1 Chemicals, reagents and media

Oligonucleotides were purchased from the NIMR oligonucleotide synthesis facility or Oswel DNA Service, Medical and Biological Sciences Building, University of Southampton. Unlabelled nucleotides were obtained from Pharmacia, and radiolabelled nucleotides from Amersham, C14 chloramphenicol was also supplied by Amersham. All other chemicals were supplied by Fisons, Sigma or BDH. Restriction and modifying enzymes were obtained from New England Biolabs, Boehringer Mannheim and International Biotechnologies Inc., except T7, T3 and SP6 RNA polymerases which were purchased from Promega and T7 DNA polymerase which was supplied by Pharmacia. Complete Protease Inhibitors Cocktail Tablets were supplied by Boehringer Mannheim.

2.2 PCR screen

Approximately 5µg of total RNA from embryo heart field (stage 22-24 tailbud embryos) and adult heart tissue were used for first strand cDNA synthesis using an oligo dT-adapter primer which contains a Xho1 site:

GActcgagTCGACATCGAT (17)

This first strand cDNA was used for PCR, using a 5' degenerate oligo (5' MSH2), which corresponded to a divergent helix 3 of an NK type homeodomain designed on the basis of sequence alignments from all known NK class genes, and a 3' nested primer:

MSH2 primer:

ACtctagaCCCAGGTGAAGATCTGGTT(C/T)CA (contains Xba 1 site)

3' nested adapter primer:

GActcgagTCGACATCG (contains Xho 1 site).

The PCR reaction used 2-5% of the cDNA, 1X PCR buffer (50mM KCl, 10mM Tris-HCl, 2.5mM $MgCl_2$, 0.2mM dATP, dGTP, dCTP and dTTP, 0.17mg/ml BSA) and 0.5µM primers. The program was as follows;

95°C 5 mins, 72°C, pause, add 1 unit Taq (1µl), 60°C 5 mins, 72°C 10 mins (1 cycle), 94°C 1 min, 60°C 1min, 72°C 2 mins 30 secs (30 cycles), 72°C 5mins (1 cycle)

The PCR product was phenol extracted, ethanol precipitated, resuspended and then digested with Xho1 and Xba1. The digest was cloned into M13 mp19 Sal1/Xba1 vector. The plaques were screened by T-Tracking to identify correct priming by the 5' MSH2 oligo. Positives were fully sequenced and screened against the BLAST databases. One clone (XNK11) showed similarity to *Drosophila bagpipe*. An internal HindIII site was used to cut XNK11 into two fragments a 5' Xba 1/Hind III fragment and a 3' Hind III fragment. Both were subcloned into pKS+ and the Xba 1/Hind III fragment was used to produce a probe for screening cDNA libraries.

2.3 Library screens

Both a *Xenopus* adult heart cDNA library (Chambers *et al,* 1994) and a stage 34-36 *Xenopus* cDNA library, were screened in order to isolate a clone containing more coding sequence. To produce the Stage 34-36 library cDNA was synthesised from stage 34-36 Poly A^+ RNA using random priming and the products cloned via EcoR1 adapters into Lambda ZapII. The libraries were lifted onto Hybond-N, Nylon, 0.45 micron membranes (Amersham). Probe was prepared using the Promega Megaprime Labelling Kit and purified using IBI Nu-clean R50 disposable spin columns (Eastman Kodak Company). Hybridisation followed standard procedure in which the pre-hybridisation (5X SSC, 5X Denharts, 0.1% Sodium pyrophosphate, 1% sarcosyl, 100mg/ml salmon sperm DNA) was incubated at 65°C for at least 1hr, with hybridisation (as prehyb but 10% dextran sulphate) being carried at 65°C overnight. The filters were briefly washed with 2X SSC at room temperature before a 1X SSC wash at 42°C. The positives were plaque purified before being excised using the *in vivo* excision EXassist/SOLR system detailed in the Stratagene ZAP-cDNA synthesis kit.

2.4 Ligation and transformation

DNA fragments were purified with the QIAquick Gel Extraction Kit (QIAGEN). Fragments were ligated into vectors with T4 DNA Ligase (New England Biolabs) using the manufacturers instruction. Reactions were incubated at RT for 1-4 hours. DNA

constructs were introduced into bacterial cells as described by Hanahan (Hanahan, 1985). Bacterial colonies containing recombinant plasmids were selected using ampicillin antibiotic selection.

2.5 Isolation of plasmid DNA

Plasmid clones were propagated in DH5α cells grown in 2XTY broth containing ampicillin. DNA from clones was isolated by the CTAB boiling method for analysis (Del Sal *et al.*, 1988). Large scale preparations of plasmid DNA were purified either by alkaline lysis and precipitation with polyethylene glycol, by equilibrium centrifugation in CsCl-Ethidium Bromide gradients (Sambrook *et al* , 1989) or by using the Hybaid Recovery, Quick Flow Maxi Kit, (Hybaid, Middlesex).

2.6 DNA sequencing

Double strand sequencing used the dideoxy method as described by Zhang *et al.*, 1988, except the denaturing step was carried out at 65°C for 15 minutes to hydrolyse any residual RNA. Nucleotide sequence was determined with the T7 DNA polymerase and α-^{35}S dCTP (10mCi/ml).

M13 cloning, transformation and DNA preparations used standard methods as described by Sambrook *et al,* 1989, sequencing as described by Sanger *et al*, 1977.

2.7 Sequence analysis

Sequences were read using a digitizer (Science Accessories corporation) and the Digigel program. The data accumulated from sequencing was compiled using the DB programs (Staden, 1982). Each consensus sequence was analysed using the ANALSEQ (Staden, 1984) and UWGCG programs (Deveraux *et al,* 1984).

2.8 Southern blot

Southern blots were carried out using standard procedure as described by Sambrook *et al*, 1989. DNA was transferred onto Hybond N (Amersham) by capillary transfer.

2.9 PCR amplification of XBap

In order to produce a myc tagged subclone that could be used in binding site selection and for the production of synthetic RNA for injection into oocytes, a PCR fragment of the whole XBap sequence was prepared using the conditions and primers described below. The product was digested with Nco I and cloned into a pβTAG vector to introduce a myc tag at the amino terminus(Fig. 4). The enzyme used was AmpliTaq DNA Polymerase (Perkin Elmer) and primers of sequence:

CCATGGCAGCTGTGCGAAGTTCC XBapORF (T)

AGCGCCATGGCAGGCAGGGCACAA XBapORF (B)

PCR conditions used 1X PCR buffer and 2μM of each primer, the program was as follows; 96° 2 mins, 55° 30 secs, 72° 40 secs, (1 cycle), 94° 40 secs, 55° 30 secs, 72° 40 secs, (20 cycles), 72° 10 mins (1 cycle).

2.10 *In vitro* transcription and translation

Capped, synthetic mRNA was transcribed from linearised template using the appropriate RNA polymerase, by standard methods (Krieg and Melton, 1984). RNA was translated using the Rabbit Reticulolysate System (Promega), an aliquot of each reaction was labelled with α-^{35}S Methionine in order to monitor the efficiency of the reaction. The products were separated on an SDS containing 10% polyacrylamide gel. SDS-PAGE was performed with a stacking gel (pH-6.8) overlying a resolving gel (pH-8.8) (Laemmli, U.K, 1970). Samples were prepared in 1X SDS denaturing buffer (50mM Tris.Cl pH6.8, 100mM DTt, 2% SDS, o.1% bromophenol blue, 10% glycerol) and heated to 95°C for 2 minutes before loading. Gels were assembled and run on the Bio-Rad Protein mini-gel system at 200V.

2.11 Immunoprecipitation

2.11.1 Method used during binding site selection.

Each reaction contained 1µl of translation (or water as control), 1µl of 9E10 antibody and 1µl of IgG (or no antibody control) 8ng/µl PolydIdC , 20µl buffer E (Dignams buffer D (see below) with the addition of protease inhibitors, 0.1% NP40, and 50µg/ml acetylated Bovine Serum Albumin) in a reaction volume of 25µl. The reaction was incubated for 30 minutes on ice. Protein A sepharose beads, 25µl per reaction, were washed in Dignams buffer D (20mM HEPES pH-7.9, 100mM KCL, 0.2 mM EDTA, 0.2mM EGTA, 20% glycerol, protease inhibitors DTT, Benzamidine and PMSF). Each reaction was added to an aliquot of the protein-A-sepharose beads and rotated at 4°C for a minimum of 1 hour. Four 750µl washes with buffer D + 0.1% NP40 were carried out followed by elution with 20µl buffer E minus NP40, and 20µl denaturing buffer. The samples were boiled for 5 minutes then 10µl of the products were separated on a 10% SDS-polyacrylamide gel as described previously.

2.11.2 Increased stringency method.

The binding reactions were carried out as above except that no IgG was added as protein-G-sepharose (Pharmacia Biotech) was used in place of protein A sepharose. After binding the beads were washed for 5 minutes each time, alternating between RIPA Buffer (150mM NaCl, 1% Np40, 0.5% Na-deoxycholate, 0.1% SDS and 50mM Tris pH-8, 2mM DTT and complete protease inhibitors), and WEEB (50mM Tris pH-8, 25% glycerol, 50mM Kcl, 0.1mM EDTA, 2mM DTT and complete protease inhibitors). In one case a final distilled water wash was included. All the reactions were carried out in silicanised tubes and before elution the reactions were transferred to a new tube.

2.12 Binding site selection

This procedure was carried out using the method as described by Pollock and Treisman, (1991), using *in vitro* translated material. This technique selected sequences bound with high affinity by XBap. Firstly a random pool of oligomers labeled with P^{32} was incubated

with XBap protein (the binding conditions as for EMSA detailed below) and then the protein/DNA complexes were immunoprecipitated. The bound oligonucleotides were eluted from the protein in 400µl of elution buffer (0.5M NH4Oac, 10mM MgAc, 1mM EDTA, 0.1% SDS) at 45°C for 3-4 hours and amplified by PCR for use in the next round of binding, immunoprecipitation and amplification.(PCR cycling program; 94°C 1 minute, 62°C 1 minute, 72°C 1 minute, for 18 cycles, one final cycle of 72°C for 5 minutes). Four rounds of selection were completed in this manner. At this stage a bandshift reaction was carried out using the probes produced by the first four rounds. The retarded band was excised from the bandshift gel and the DNA eluted. After phenol extraction and ethanol precipitation the DNA was resuspended in 10µl of which 3µl were used in the standard PCR reaction. Three more rounds of selection were carried out before the PCR products from the final round of selection were digested with Bam H1 and Eco R1 and cloned into an mp18 Eco R1/ Bam H1 vector.

2.13 Electrophoretic mobility shift assay (EMSA)

Binding reactions contained 1µl of *in vitro* translated or purified protein in a reaction volume of 12µl under buffer conditions which had been shown to be optimal for DNA binding of TTF-1 protein (Lazzaro *et al*, 1991). (2X Binding Buffer:20mM Tris-HCl pH7.6, 75mM KCl, 0.25 mg/ml BSA, 1mM DTT and 10% glycerol. PolydIdC was added to the reactions to a final concentration of 83ug/µl). Binding reactions were incubated at RT (must be below 23°C) for 15 minutes before the addition of 20,000cpm of probe and a further 15 minute RT incubation. Antibody shift analysis involved the addition of 1µl of 9E10 antibody and an extra 15 minutes incubation. Complexes were resolved on a 6% polyacrylamide gel in 0.5X TBE buffer and ran at 100-200V at 4°C. Gels were dried unfixed onto 3MM paper for autoradiography. The oligonucleotides used for EMSA were annealed by incubating for 10 minutes at 88°C followed by slow cooling to RT. Labeling was achieved by 3' fill-in with p^{32}-dCTP (3,000 Ci/mmol) using the Klenow fragment (Promega).

2.14 pGex bacterial expression system

Appropriate DNA constructs were transformed into the protease deficient Epicurian Coli BL21 (DE3) Competent cells, (Stratagene). An overnight culture was diluted 1/10 and grown at 37°C for 1 hour. IPTG was added to a final concentration of 0.1mM and the culture grown for a further 5 hours. At the end of the incubation period the cells were pelleted by centrifugation before resuspension in 1ml MTPBS per 100 mls culture (2mM EDTA, 2mM orthophenanthrolene, 1X complete protease inhibitors (Boehringer Mannheim)). The cells were lysed by passing through a French press (1000lbs/sqin) and Triton X-100 added to 1%. After centrifugation at 6000rpm for 15 minutes the supernatant was added to half volume 50% glutathionine agarose beads and rocked at RT for 2 minutes. The beads were washed three times with 5 volumes of MTPBS + 1% Triton X-100 and two times with 5 volumes of MTPBS. Protein was eluted from the beads by 2 minutes RT rocking with 5mM fresh reduced glutathione. After a quick spin the supernatant was removed, after a second elution step the supernatants were pooled.

2.15 Site directed mutagenesis

The mutant form of XBap protein was generated using the QuikChange Site-Directed Mutagenesis Kit (Stratagene) with the following primers:

GTCTGGCCCTGAGCGAAGCGACATGGCTGCCTCC XBap SDM (T)

GGAGGCAGCCATGTCGCTTCGCTCAGGGCCAGAC XBap SDM (B)

This kit employs a PCR based strategy requiring the design of two PCR primers containing the desired mutation (Figure X.C). The XBap construct ORF/Myc/pT7 was used as template in the PCR reaction with the SDM primers and *pfu* DNA polymerase. Following the cycling reactions the mixture was digested with Dpn 1 enzyme which selectively digests the methylated, non-mutated parental DNA template. The circular, nicked dsDNA is then transformed into XL1 competent cells where the nicks are repaired.

2.16 Oocyte assay system

The XBap Nco PCR pβTAG subclone was digested with Eco R1 and Hind III to produce a fragment containing the XBap sequence plus the myc tag. This fragment was subcloned into pT7TS(R1-) in order that synthetic RNA of the XBap ORF could be produced for injection into oocytes.

Adult *Xenopus* ovaries were teased into small fragments and collagenase (Sigma) treated for 1-2 hrs. Calcium and magnesium free OR2 (OR2-, NaCl 82.5mM, KCl 2.5mM, Na_2HPO_4 1mM, HEPES 5mM, PVP 0.5g/l) medium was used for washing oocytes to remove collagenase and oocytes were maintained in OR2+ medium (OR2-, supplemented with 1mM Ca and Mg chlorides). All injections were carried out in 2% Ficoll in OR2+. Approximately 5-10ng of RNA was injected into the vegetal pole and the oocytes incubated at 20°C for 12-24 hours before injection of 5ng of DNA into the germinal vesicle. The pBLCAT3T constructs injected are illustrated in Fig.8. After incubation for 24 hours the oocytes are harvested by freezing on dry ice.

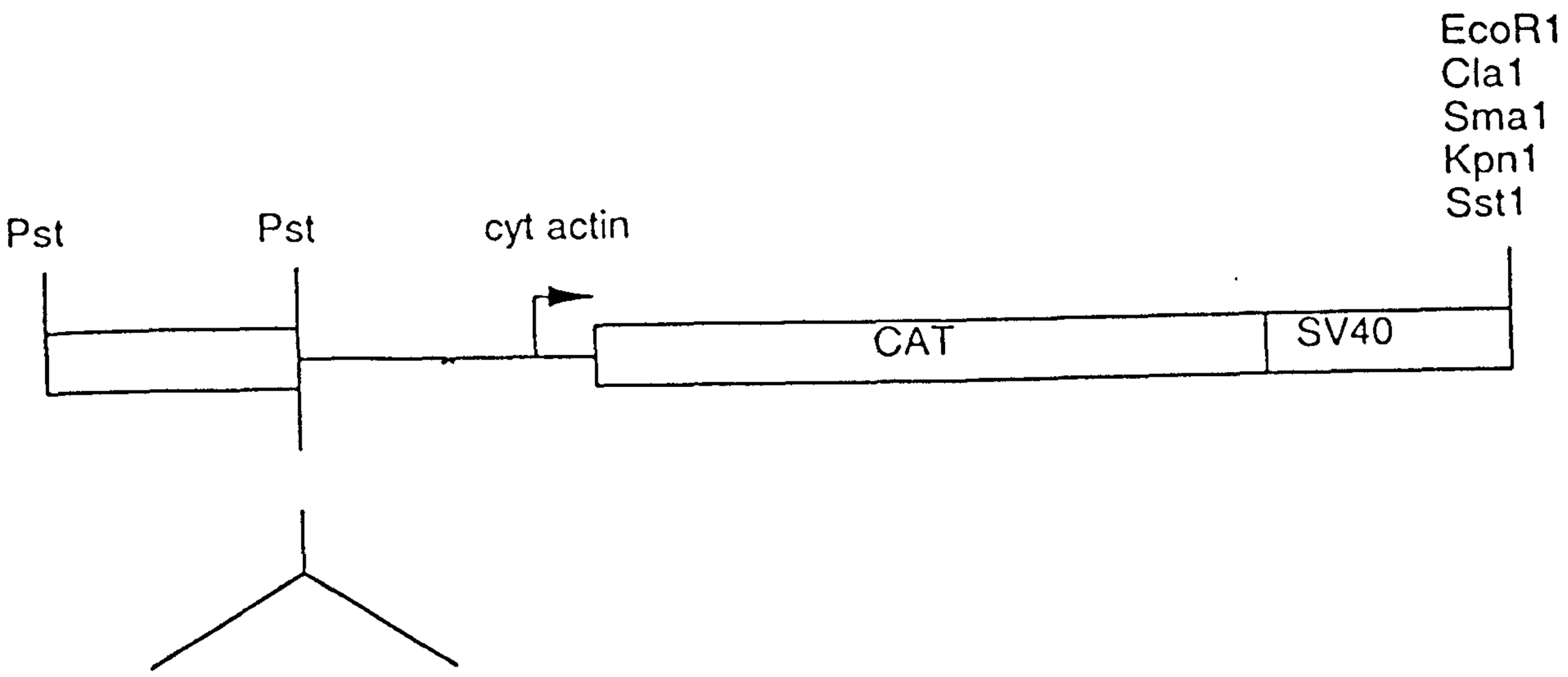
In each experiment approximately 100 oocytes were injected for every DNA/RNA combination. Each of these groups were divided into three pools for CAT assay. CAT assays were carried out according to Etkin *et al,* 1985, with the modification that samples were added to a master mix containing the chloramphenicol and acetyl CoA and the mixture incubated for precisely 1 hour. Reactions were separated by thin layer chromatography on Polygram SIL G/UV for TLC, 0.25 mm silica gel. Chloramphenicol conversion to acetylated forms was quantitated by cutting out the areas on the silica gels and determining the radioactive counts in a scintillation counter. Reactions were repeated where necessary using reduced amounts of extract to ensure CAT activity remained within linear range.

2.17 Western blotting of oocyte extract

Following overnight incubation 100 injected oocytes were homogenised in 150µl of buffer A (0.05M Tris pH8, 40% glycerol, 100mM KCL, 5mM Mg, 0.1% b-Mercatoethanol, complete protease inhibitors). The samples were spun at 14, 000 rpm for 30 minutes at 4°C. The supernatant was removed and spun at 100, 000 rpm for 30 minutes at 4°C. To the supernatant 20µl 10% SDS, 4µl 0.5M EDTA and 2µl 100mM DTT was added. Samples were prepared with 2X sample buffer and loaded onto a 10% SDS-polyacrylamide gel. The pellet was resuspended in 32µl of buffer A made to 0.3M NaCl and 1mg/ml DNase and incubated at 37°C for 10 minutes. This reaction was then made to 10mM EDTA and 1% SDS before samples were prepared and loaded onto a 10% SDS-polyacrylamide gel. After running the gel at 200V for 45 minutes the gel was soaked in CAPS buffer (1X CAPS, 10% Methanol) for 15 minutes along with the membrane (pre-wetted in methanol) (Immobilon P, Millipore), sponges and sheets of blotting paper needed to make up a transfer sandwich. The western blot was carried out at 50V/170mA for 30 minutes using equipment supplied by BioRad.

Following transfer the membrane was blocked with 5% milk powder in PBS at RT for 1 hour. 9E10 antibody (Santa Cruse) diluted to 3mg/ml was added to a 0.5% milk solution and the membrane incubated at RT for a further 90 minutes. Four 15 minute RT washes were then carried out with 0.1% Tween/PBS. The secondary antibody, (antibody-mouse IgG Alkaline Phosphatase conjugate, Sigma Immuno chemicals), was diluted 1/1000 and incubated with 0.5% milk/PBS for 90 minutes. Four 15 minute RT washes with 0.1% Tween/PBS were followed by a 15 minute wash in western developing buffer (100mM Tris pH-9.5, 100mMNaCl, 5mM $MgCl_2$). The membrane was developed in 20µl 5-Bromo-4-chloro-3-indoyl-phosphate 50mg/ml, 20µl Nitro Blue Tetrazolium 75 mg/ml, 70% Di-methyl formamide made upto 50mls with developing buffer. The blot was developed in the dark for 15 minutes after which the reaction was terminated with stop buffer (100mM Tris, 100mM NaCl, 10 mM EDTA).

Figure. 8 Diagram of pBLCAT3T reporter constructs used in *Xenopus* oocyte transcription assays.



oligos

TTf-1	ctagCCCAGTCAAGTGTTCT	1, 3 or 5 copies
TTF-1mut	ctagCCCAGTCTCATGTTCT	5 copies
A20	ctagCCCAGTTAATTGTTCT	1 or 4 copies
Xbap	ctagGATTTAAGTGGTCGTTAAGTGGTCG	3 copies

2.18 Mammalian cell transfection

Mammalian cell transfection were carried out with NIH 3T3 fibroblasts (embryo, contact-inhibited, NIH swiss mouse, American Type Culture Collection catalogue number ATCC CRL 1658). Cells were cultured in Dulbeco's modified Eagles medium with 4.5g/L glucose, 90%; calf serum, 10%. Lipofectamine transfection was used in experiments to study localisation of protein and calcium phosphate transfection was used in transcriptional activation studies. pRL-SV40 or pSV-β-galactosidase (both Promega UK Ltd, Southampton) were used as controls of transfection efficiency.

2.19 Lipofectamine cell transfection

Cells were plated out into 6 well tissue culture dishes with four cover slips in each well, at 1.25 x 10^5 cells/well, the day before transfection, and incubated overnight at 37°C/5% CO^2. DNA to be transfected was diluted with PBS such that the final concentration of DNA per transfection was 0.4 - 0.6µg. 4-6µl of lipofectamine (GibcoBRL, 2mg/ml) was added to 10-12µl of PBS (10:1 lipo:DNA). To these droplets 16µl of diluted DNA was added and mixed thoroughly. The reactions were incubated at RT for 30 minutes. The cells were washed twice with serum free medium and after the final wash 800µl of serum free medium was added to each well. At the end of the incubation 200µl of serum free medium was added to each DNA-:Lipofectamine sample and the whole sample was sprinkled over a well of cells which were then incubated for six hours. Following incubation the cells were washed twice in serum containing medium, with 2.5mls of medium being added to each well after the final wash. The cells were harvested for analysis following incubation for 48hours.

2.20 Immunocytochemistry

Cells were fixed in 3% paraformaldehyde in PBS for 30 minutes. Each cover slip was then washed with PBS 0.1% Tween for 5 minutes. Blocking was carried out by covering each slip with PBS 0.25% Fish Gelatin (spare cover slips were stored in PBS 0.2% Azide at 4). Anti HA antibody (12CAG) was diluted 1/3000 and anti-myc (9E10) 1/156 with

PBS 0.25 % gelatin. Each cover slip was covered with 200µl of antibody mix and incubated at 4°C overnight. The following day the slips were washed thoroughly in PBS before blocking for 20 minutes with PBS 0.25 % gelatin. Flourescin (FITC)-conjugated affinity pure goat anti-mouse IgG was added and incubated at RT for one hour. The cells were then washed in PBS before mounting with a drop of Citiflour (UKC, Chem.Lab).

2.21 Calcium phosphate transfection

Calcium phosphate transfection was carried out using standard protocols as described by Manniatis *et al.* A total of 5µg DNA was transfected per well. The proportions of DNA transfected were 70% test plasmid, 10% reference construct (pSV-β-Galactosidase Control Vector, Promega), 20% reporter construct (pBLCAT3T).

2.22 Preparation of transfected cell extract

Cells were harvested by washing twice with PBS before scraping cells from each well into 1ml of 0.5mM EDTA/PBS. This cell suspension was spun at 13, 000K for 5 minutes, and the pellet resuspended in 200µl of Reporter Lysis Buffer (Promega). The lysed cells were spun at 13, 000K to pellet the debris. This extract was then assayed for β-galactosidase activity (Sambrook *et al*, 1989) and CAT activity. The CAT assay followed the protocol described earlier with the modification that 40µl of extract was used and the reactions were incubated for 2 hours.

2.23 Luciferase assay

Luciferase activity of transfected cell extracts was monitored using the Dual-Luciferase Reporter Assay System (Promega UK Ltd, Southampton). Luminescence was measured in a CliniLumat, Berthold, Luminometer.

2.24 DNA constructs

Orig/pKS+ (1)

The original construct plaque purified from the stage 34-36 cDNA library produced an Eco RI fragment of 1.4Kb, containing the XBap open reading frame, in a pKS+ vector (Stratagene Ltd, Cambridge, UK).

(XBap 1-1400; 1-528 5'UTR, 528-1350 ORF, 1351-1400 3'UTR. TN domain 558-588, HD 966-1146, NK2 domain 1173-1218. All construct co-ordinates are given as fragment of this construct, (see Fig.9.)

Orig/pT7 (2)

Construct 1 was digested with Eco RI and the 1.4kb fragment subcloned into pT7TS(R1+) (Fig.10). This subclone produces sense RNA when linearised with Xba I and transcribed with T7 RNA Polymerase.

(XBap 1-1400; 1-528 5'UTR, 528-1350 ORF, 1351-1400 3'UTR.)

ORF/Myc/pβ (3)

The open reading frame of XBap was amplified by PCR from construct 1 using primers (XBapORFTop and XBapORFBot) each of which contain an Nco I site. The PCR product was purified and digested with Nco I. The resulting fragment was cloned into a pβTAG Nco I vector (Pollock and Treisman, 1991 B). This cloning introduces a myc tag to the N-Terminus of the XBap open reading frame.

(XBap 529-1311, the complete ORF)

ORF/Myc/pT7 (4)

Construct 3 was digested with Hind III and Eco RI and subcloned into a pT7TS(R1+) vector. This cloning correctly maintains the XBap reading frame. Linearisation with Xba I and transcription with T7 RNA Polymerase produces sense RNA. Translation of this RNA produces XBap protein with a myc tag.

(XBap 529-1311, the complete ORF)

ORF/Myc/pC (5)

Construct 4 was digested with Xba I and Hind III to produce a fragment containing both the XBap open reading frame and the myc tag. This fragment was cloned into a pCDNA3 vector (Invitrogen, The Netherlands) which correctly maintained the reading frame of XBap. This construct is suitable for mammalian cell transfection.

(XBap 529-1311, the complete ORF)

ORF/Ha/pT7 (6)

Construct 3 was digested with Nco I and cloned into a pT7TS-HA vector (Fig.10). Linearisation with Xba and transcription with T7 RNA polymerase produces sense RNA. Translation of this RNA produces the full open reading frame of XBap with an HA tag, which is the epitope of the 12CAG antibody, at the amino terminus.

(XBap 529-1311, the complete ORF)

ORF/Ha/pC (7)

Construct 6 was digested with Xba I and Hind III to produce a fragment containing the open reading frame and the HA Tag. This fragment was cloned into pCDNA3 (Invitrogen, The Netherlands) which is suitable for mammalian cell transfection.

(XBap 529-1311, the complete ORF)

HD/Ha/pT7 (8)

The homeobox of XBap was amplified by PCR using primers X and Y from construct 1. Following purification of the fragment it was digested with Nco I and cloned into a pT7TS (Nco) HA vector. Linearisation of this template with Xba I and transcription with T7 RNA Polymerase produces sense RNA. Translation of this RNA produces the homeodomain of XBap with an HA Tag.

(XBap 962-511, XBap homeodomain)

HD/Ha/pC (9)

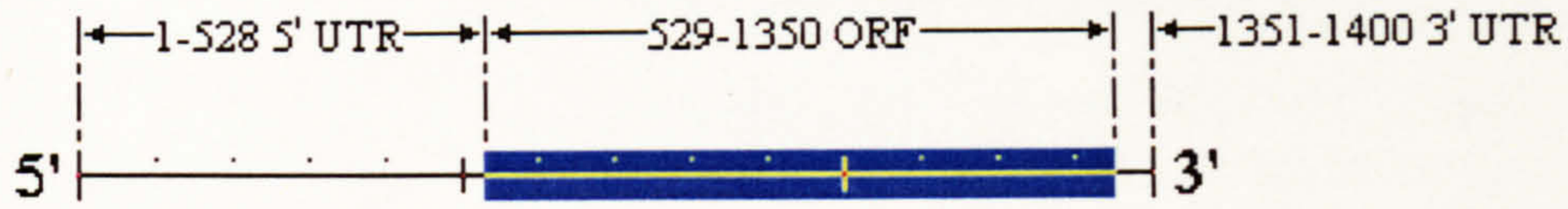
Construct 8 was digested with Xba I and Hind III and the fragment, containing the homeodomain and the HA Tag, was cloned into pCDNA3 (Invitrogen, The Netherlands).

Figure 9. Schematic representation of cloning 1 (orig/pKS+) on which all other clonings are based

A To scale representation of full insert of cloning orig/pKS+

B To scale representation of the open reading frame of XBap, a fragment of the orig/pKS+ insert used in many subsequent clonings.

A


1-528 5' UTR
529-1350 ORF
1351-1400 3' UTR
5'
3'

B

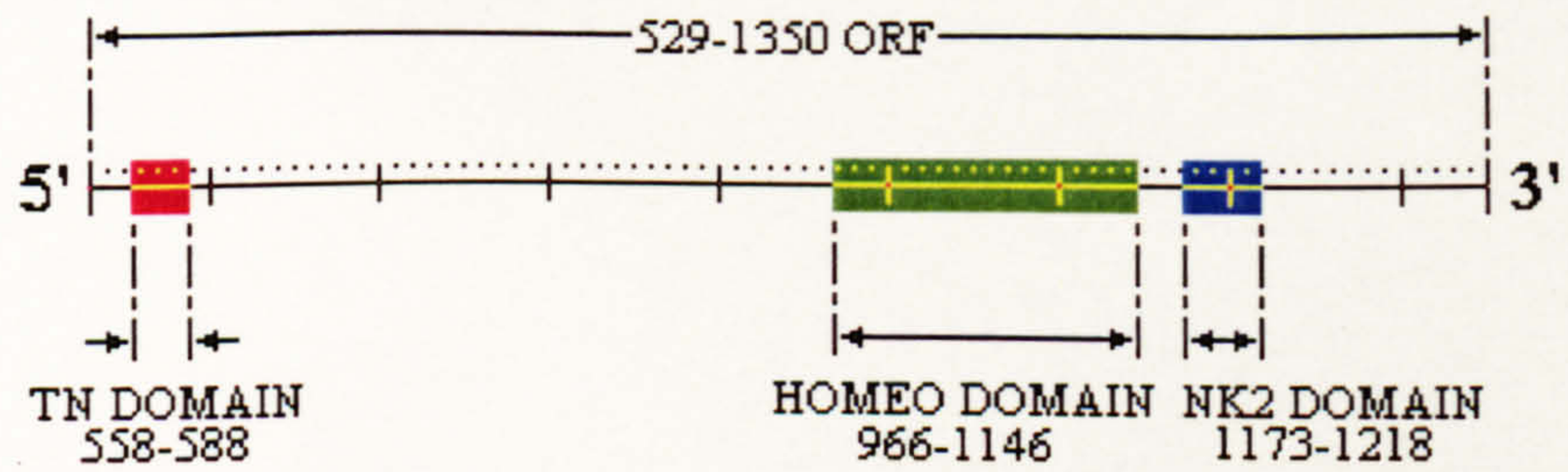
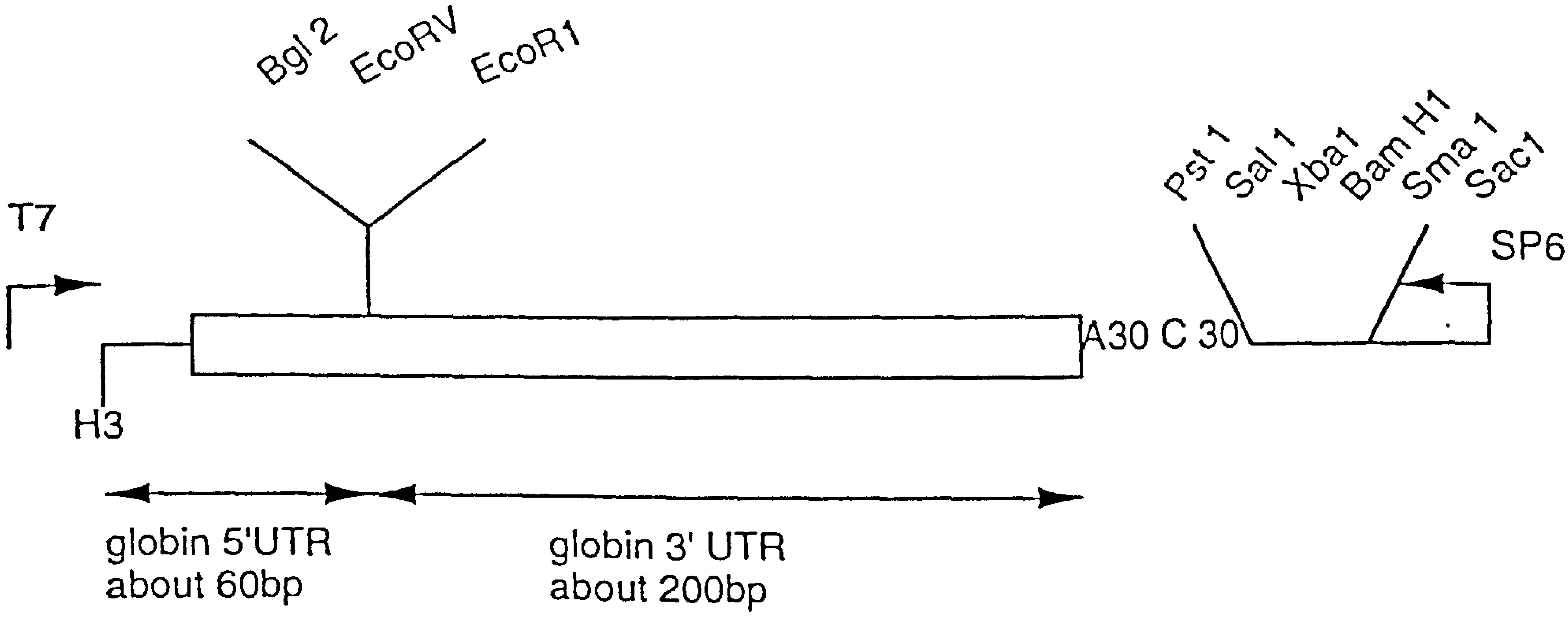

529-1350 ORF
5'
3'
TN DOMAIN
558-588
HOMEO DOMAIN
966-1146
NK2 DOMAIN
1173-1218

Figure 10. Transcription plasmid T7TS R1-

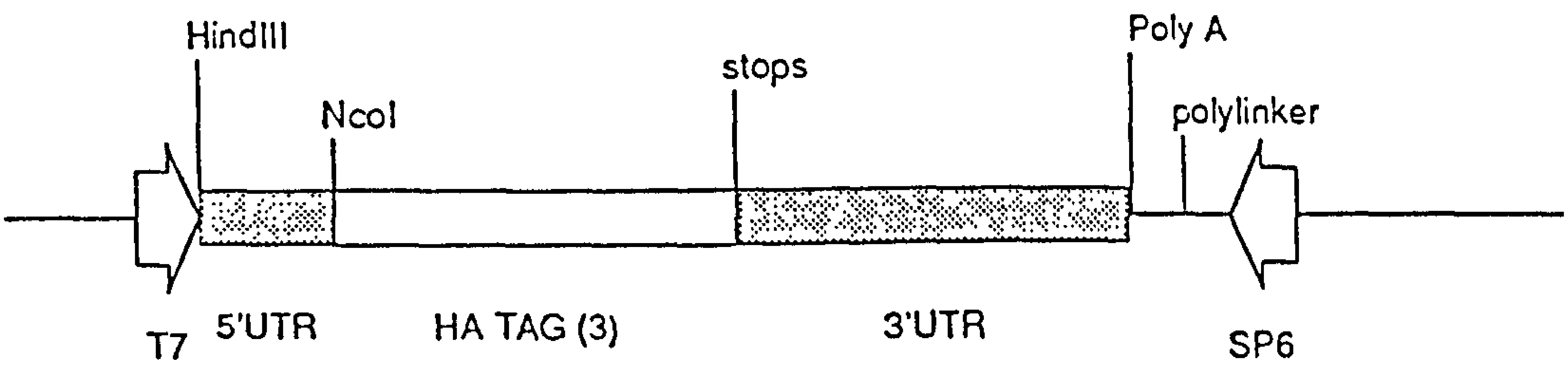
5' and 3' untranslated regions of *Xenopus* globin mRNA inserted into pGEM4Z. Sense strand transcription driven by T7 promoter. Vector has stop codons in all three reading frames downstream of multiple cloning site.

PT7TS.HA(NCO), modification of pT7TSR1- which introduces three copies of the HA tag.



Sequence of 5' and 3' UTRs and polylinker:
AAGCTTGCTTGTTCTTTTTGCAGAAGCTCAGAATAAACGCTCAACTTTGGCACAGAT
CTGATATCGAATTCGGATCTGAGTGACTGACTAGGATCTGGTTACCACTAAACCAG
CCTCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCAACTTACACTTACAA
AATGTTGTCCCCCAAAATGTAGCCATTCGTATCTGCTCCTAATAAAAAGAAAGTTTC
TTCACATTCTA30 C30

pT7TS.HA(Nco)



AAGCTTGCTTGTTCTTTTTGCAGAAGCTCAGAATAAACGCTCAACTTTGG
CAgccatggTTTACCCATACGATGTTCCTGACTATGCGGGCTATCCCTAT
GACGTCCCGGACTATGCAGGATCCTATCCATATGACGTTCCAGATTACGC
TGCTCAGTGCGGCCGATCTGCTAGTGACTGACTAGGATCTGGTTACCACT
AAACCAGCCTCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCA
ACTTACACTTACAAAATGTTGTCCCCCAAAATGTAGCCATTCGTATCTGC
TCCTAATAAAAAGAAAGTTTCTTCACATTCTAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCtgcaggtcga
ctctagaggatccccgggtaccgagctcgaattc

(XBap 962-511, XBap homeodomain)

C-Del/pT7 (10)

Construct 6 was used in a PCR amplification with T7 primer and 'HD bottom'. The product was digested with Nco and cloned into a pT7TS(R1+). Linearisation with Xba I and transcription with T7 RNA Polymerase produces sense RNA. Translation of this RNA produces XBap protein that has all amino acids C-Terminal of the homeodomain, including the NK2 domain, deleted.

(XBap 529-1156, XBap TN domain and homeodomain included)

N-Del/pSP (11)

Construct 2 was linearised with Bam HI and the overhang was filled in using klenow fragment. The product was digested with Eco RI to produce a blunt/Eco R1 fragment. A vector was prepared by linearising pSP64.T (Promega UK Ltd, Southampton) with Nco I, this was filled in using klenow fragment. This DNA was then cut with Eco R1 producing a blunt/Eco R1 pSP64.T vector into which the XBap fragment was cloned. Linearisation of this construct with Eco RI and transcription with SP6 RNA Polymerase produces sense RNA. Translation of this RNA produces XBap protein that has all amino acids N-terminal to the homeodomain, including the TN domain, is deleted.

(XBap 922-1400, XBap homeodomain and NK2 box included)

DBap/pG (12)

A clone containing the homeodomain of *Drosophila Bagpipe* as a Pst I fragment in a PQE vector was kindly supplied by Dr M. Frasch. This clone was digested with Bam HI and Pvu II and cloned into a pGEX2T Bam HI / Sma I vector (Promega UK. Ltd, Southampton). Protein was prepared using IPTG stimulation and extraction of protein from the bacterial culture.

(Dro Bap ORF 523-990, homeodomain 634-834)

Bx1/pSP (13)

A clone containing the full open reading frame of mouse *Bapx1* as a 1.0Kb Bam H1 / Hind III fragment in a pQE9 vector was kindly supplied by Dr. D. Neustaedter. This clone was digested with Bam HI and Hind III and the fragment was cloned into pSP65 (Promega UK Ltd, Southampton). Linearisation with Hind III and transcription with SP6 RNA Polymerase produces sense RNA.

(*Bapx1* full open reading frame)

N3.1/pC (14)

A clone encoding amino acids 1-237 of Nkx-3.1 as a Kpn I/ Xba I fragment in a pCDNA3 vector (Invitrogen, The Netherlands) was kindly supplied by Dr. Abate-Chen. Linearisation with Xba I and transcription with T7 RNA Polymerase produces sense RNA.

(Nkx-3.1, complete ORF)

XBF/BGal (15)

Construct 4 was used in a PCR reaction with SP6 primer and XBap R1 primer. The PCR fragment was digested with Eco R1 and Xba 1 and cloned into an Eco R1/Xba 1 pGBX1 vector (kindly supplied by Dr Joshua Brickman, unpublished). On tranfection into ES cells this construct produces full length XBap protein fused to the DNA binding domain of Gal4.

(XBap 529-1311, the complete ORF)

XBCD/BGal (16)

Construct 10 was used in a PCR reaction with SP6 primer and XBap R1 primer. The PCR fragment was digested with Eco R1 and Xba 1 and cloned into an Eco R1/Xba 1 pGBX1 vector (kindly supplied by Dr J. Brickman, unpublished). On transfection into ES cells this construct produces a protein in which the N-terminal region and homeodomain of XBap is fused to the DNA binding domain of Gal4.

(XBap 529-1156, XBap TN domain and homeodomain included)

2.25 Synthetic oligos

2.25.1 PCR primers

ACtctagaCCCAGGTGAAGATCTGGTT(C/T)CA	MSH2		
GActcgagTCGACATCG	3' nested primer		
CCATGGCAGCTGTGCGAAGTTCC	XBapORF (T)	23	Nco1
AGCGCCATGGCAGGCAGGGCACAA	XBapORF (B)	24	Nco 1
CCCCCATGGAGAGAAAGAAGCGCTC	XBapHD(T)	25	Nco1
AAGAACCATGGTAGCCATCTGCCTG	XBapHD (B)	25	Nco1

2.25.2 Binding oligos

CTAGGATTTAAGTGGTCGTTAAGTGGTCG	XBap T2 (T)	29	Xba1
CTAGCGACCACTTAACGACCACTTAAATC	XBap T2 (B)	29	Xba1
CTAGGATTTAAGTGGTCGTTCGATGGTCG	XBap SMut (T)	29	Xba1
CTAGCGACCATCGAACGACCACTTAAATC	XBap SMut (B)	29	Xba1
CTAGGATTTAAGTGGTCGGCCGACAATCG	XBap SMut2 (T)	29	
CTAGCGATTGTCGGCCGACCACTTAAATC	XBap SMut2 (B)	29	
CTAGCGACCATCGAACGACCATGCAAATC	XBap DMut (T)	29	Xba1
CTAGGATTTGCATGGTCGTTCGATGGTCG	XBap DMut (B)	29	Xba1
CTAGGATTTAAGTGGTCGTcAAGTGGTCG	XBap C1 (T)	29	Xba1
CTAGCGACCACTTgACGACCACTTAAATC	XBap C1 (B)	29	Xba1
CTAGGATTcAAGTGGTCGTcAAGTGGTCG	XBap C2 (T)	29	Xba1
CTAGCGACCACTTgACGACCACTTgAATC	XBap C2 (B)	29	Xba1
CTAGAGGGTTATTTTTAGAGCG	NS (B)	23	
CTAGCGCTCTAAAAATAACCCT	NS (T)	23	

2.25.3 SDM primers

GTCTGGCCCTGAGCGAAGCGACATGGCTGCCTCC	XBap SDM (T)	34
GGAGGCAGCCATGTCGCTTCGCTCAGGGCCAGAC	XBap SDM (B)	34

Chapter 3

Isolation of *Xenopus Bagpipe*

3.0 Chapter 3: Isolation of *Xenopus Bagpipe*

3.1 Introduction

At present very little is known about the subdivision of the vertebrate mesoderm into cardiac, somatic and visceral mesoderm. In *Drosophila* however, studies of the expression and of null mutations of the gene *tinman* (NK4), have shown that it is essential for formation of the dorsal vessel (heart) and visceral mesoderm. Genetic evidence suggests that *bagpipe* (NK3) is a target of *tinman*, and mutations in this gene result in failure of visceral mesoderm to differentiate into midgut musculature. Thus *tinman* and *bagpipe*, in association with other factors, are good candidates for specification of visceral mesoderm in *Drosophila.* The cloning and characterisation of a *Xenopus Bagpipe* relative was, therefore, of great interest.

A PCR based strategy followed by full library screening had proven to be successful in the cloning of other vertebrate *NK* class genes (X*Nkx-2.5*, Tonissen at al, 1994). Given the co-expression of *tinman* and *bagpipe* in the *Drosophila* dorsal vessel it was decided that a screen of *Xenopus* heart cDNA may prove to be the most fruitful as both *bagpipe* and *tinman* relatives may potentially be isolated.

3.2 Results and discussion

3.2.1 Cloning of *Xenopus Bagpipe*

A PCR screen for *NK2* class related clones had been carried out previously in the laboratory by Robert Wilson (unpublished data). An alignment of all the known NK class gene had revealed that helix three of the homeodomain illustrated the highest degree of conservation between the genes and therefore it was this region which was used in the primer design. This primer (MSH2 - see materials and methods) was then used in an RT-PCR reaction with single-stranded cDNA prepared from *Xenopus* adult heart tissue using an extended oligo-dT primer and a 3' nested PCR primer. One of the 12 positives isolated (XNK11) showed significant homology to *Drosophila bagpipe* but had no homology to any other gene when screened against the BLAST database. The XNK11 insert was

isolated by an Xba1 / HindIII digest, generating two fragments due to an internal Hind III site and these were subcloned into pKS+. The 5' Xba/ Hind III subclone was used to prepare a probe for screening cDNA libraries. Full sequencing of the XNK11 subclones was (1.14Kb) and is illustrated in figure 11. The presence of a poly-A+ tail indicated that the original clone contained the 3' end of the gene.

cDNA library screens

Initially a *Xenopus* adult heart cDNA library was screened using the radio-labelled DNA Xba/Hind III probe and one positive was produced that corresponded to a region within XNK11(data not shown). In order to isolate a larger clone a random primed stage 32-6 tadpole cDNA library was screened. Approximately 1 million plaques were screened and five positives were isolated. Restriction enzyme mapping and sequencing of the 5' and 3' ends, showed these to be identical. The clone (orig/pKS+) was digested with Alu and Sau restriction enzymes and the fragments were subcloned into an M13 vector (mp19). The M13 sequence revealed further restriction enzyme sites and these were used to produce a series of subclones that enabled the assembly of contig which covered the whole clone (1.4Kb). Both strands were sequenced and in the generation of the final contig every base had been independently sequenced at least three times. The full sequence and putative open reading frame of the cDNA clone are illustrated in figure 12. This clone contains the full open reading frame and overlaps with the clone produced by the 3'RACE screen but does not itself contain the poly-A+ tail. The sequence produced by overlapping the cDNA and 3'RACE clones is illustrated in figure.13. The entire sequence includes 528bp 5'UTR, an open reading frame of 822bp and a 3'UTR of 883bp.

Genomic library screen

A genomic library screen was carried out in parallel to the cDNA library screen using the same probe and resulted in the isolation of five positives. The cDNA clone (Orig/pKS+) was used to generate two probes, one corresponding to a region upstream of the homeobox (Eco R1 / Bam H1) and one downstream (Sma 1 / Eco R1). These were used

Figure 11. Full sequence of the 3'RACE derived clone.

Underlined region indicates the match with the MSH2 primer used in the 3'RACE.

MSH2 primer: ACtctagaCCCAGGTGAAGATCTGGTTTCA	
Dro Nk2	ACCCAGGTGAAGATCTGGTTTCA
Dro Nk3	ACACAGGTCAAAATTTGGTTCCA
Dro NK4	ACCCAAGTGAAGATTTGGTTCCA

```
      |    10      |    20      |    30      |    40      |    50      |    60
    1 CCCAGGTGAA GATATGGTTC CAGAACAGGC GCTACAAGAC CAAGCGCAGG CAGATGGCTA   60
   61 CCGACCTTCT TGCTGCAGCC CCTGCAGCAA AGAAAGTTGC TGTGAAAGTG CTGGTGAGGG  120
  121 ACGATCAGAG ACAGTACCAT CCTGGGGAGG TCCTTCACCC TTCACTGCTT CCCCTTCAGG  180
  181 CCCCTTATTT CTACCCTTAT TATTGTGCCC TGCCTGGCTG GACGCTCTCC GCAGCTGCCT  240
  241 GCAGTAGGGG GACACCATAA AACAAGGGGG TAACTACTTA TGATCACCcC CcCCTTTGCA  300
  301 CCCTGGGGGC TTGAATATAG ACTCGTTTAT TCTGCACCAG CAAAACTGTG ACCAAAAAAG  360
  361 GATTTTGAAC AATTGCAGAC TTTGTCCATG GTATCCATGT TGTCTATGTG CCGTATAGGT  420
  421 TTTATATCTA TATTATAGCT CCTTATATTG TCTATGGAAT GACTAAGCTT TCTGGGCGAG  480
  481 TTGAGGGGTC TGTAAAAAAG GGGATCCCGG GTTCGGACTC TTGGTGGCCC CCACTGACTG  540
  541 CATGGGGCTG TTTCTAGGTT CCATGGGTGT CCAGGGCTAC GAATGTATAG TACAGAGCTT  600
  601 CCTTATCGCT TTACCATTCC CCATCCTGCC TCGTTCTTAC TCTCCTCATC ATCCATTAGA  660
  661 TTGACGTCAT CTAAAGGGGT CTGCATAGGT AGCAATCCAT AGAGATGTAC TACGTTAGGT  720
  721 GGGGAGGTAA GACATTTTTC TAGCACTTTT ATTCTCTTGC AACCTGTGGC CTTTCTCTTG  780
  781 CAACTATCTG TCTCTGCCTT TCCCAGACAG TTGGGAATAT AAATCCAACT ACAAATGGGG  840
  841 AACCTTTGGA TGGATAATCT GGCAGAGATA CTTGCTAGAG CTACGTGCTT TTATCTCCGC  900
  901 TCAGCTCAGT GCCTTTCCCG AAAAACACTT AGACACAGGC ATCTGCTTCA TTGTTGCCTT  960
  961 ATTTATAAAC ACAAGATGCA ATCCTTTAAT AATCCCGCAG TTTGTAGATT CACAAGAATC 1020
 1021 TGGCACTGCG GCATCAGCTG CATTACTGTA TGTGCTTTTC CAATGACATT GTGATAAGTG 1080
 1081 TTTTTTAATA CTTTGTTTTT TTTCTAATTT ATAAAGTATA GTTTAACTAA AAAAAAAAAA 1140
 1141 A                                                               1141
      |    10      |    20      |    30      |    40      |    50      |    60
```

Figure 12. Full sequence of the cDNA clone isolated from a random Stage 36 *Xenopus* cDNA library.

The underlined regions indicate the R1 adaptors.

```
          |    10      |    20      |    30      |    40      |    50      |    60
   1 GAATTCGGCA CGAGAGCGGC TCAGAGCTCA GCCTGTGGCT CCTGCACTCT GTCTGCCCTG  60
  61 AAGGCTGTCA CTGCCATTGC CTGTGCTGGG GCTGCTTACT TCTTGCTACA CTGAATACTG 120
 121 GCCCTGCAGA AGGGACACTA CCTGATCTGT TAAACTTGCA GGTTGATATT CCTACAGCTT 180
 181 GCCAGTCACA TAGACAGTCA GCCCTCGTGC CAACCCGACT TGCTGCCAAT CATGCAGTGT 240
 241 GCCTGACCCA CAAGCTGTCA CCCCAGGGAG TGGGCTACAA GCAGAGAAGA CTGTGGCTTT 300
 301 TCCTTAATGA GACCTGTGGA TCAGTAGATT TGCAAGATAC ACTGGTGGTC ACACCCTGGT 360
 361 CCCTTTATGT TCAGAGGCCC TCTGGATACC CCCTGGGCAG CAGATGTTTC CCCCAGTGGC 420
 421 CCACGCGGGA TCAAGGCTGG ACGCTCTGCC CTGCTAGTGC CCATCTTGGG ATCCATTTGG 480
 481 GGAGCTTGGC GCCCAGAGCC AGGAGCCACC CGGTACCGGA CACTGGATCC AATGGCAGCT 540
 541 GTGCGAAGTT CCAGCAGGCT GACTCCATTC TCTATCCAGG CTATTCTGAA CAGGAAAGAG 600
 601 GAGCGTGCCC ACACGTTCCC CAGGCTGGCC ACTGCAGTCT CACCCGCCTG TTGCTGGAGG 660
 661 ATCTTTGGGG AGAGTGAAGC AGAAGGACTG GGCTCCCCAT GTGGAGGGGC CCCTGGGGCA 720
 721 GGAGCCGGGG GAGAGCCATC CGGGTGGGAC TCGGATTCAG CCCTCAGTGA GGAGGGAGAA 780
 781 CTGGGAAGGC CGGGGGATCT AGGGGAGAGG AAGAAGCAGA GGCCGCTGGA GGCCAGAGCC 840
 841 AAGGGGGAGG ATGAGGAGGA AACCCCAGGC TGCAGCGACT GCGAGATAGG AGCCAGTGTC 900
 901 CCAGATCCCA GCCCTCCAGA TGAGGATCCC AAGTGTGAGC AGCTGATGCT AGAGCCCCCC 960
 961 AAGCAGAGAA AGAAGCGCTC CCGGGCTGCA TTCTCCCATG CCCAGGTCTT TGAGCTGGAG 1020
1021 AGACGCTTCA ACCACCAGCG ATACTTGTCT GGCCCTGAGC GAGCTGACCT GGCTGCCTCC 1080
1081 CTTAAACTAA CAGAGACCCA GGTGAAGATA TGGTTCCAGA ACAGGCGCTA CAAGACCAAG 1140
1141 CGCAGGCAGA TGGCTACCGA CCTTCTTGCT GCAGCCCCTG CAGCAAAGAA AGTTGCTGTG 1200
1201 AAAGTGCTGG TGAGGGACGA TCAGAGACAG TACCATCCTG GGGAGGTCCT ACCCTTCACT 1260
1261 GCtTCCCCTT CAGGCCCCTT ATTTCTACCC TTATTATTGT GCCCTGCCTG GCTGGACGCT 1320
1321 CTCCGCAGCT GCCTGCAGTA GGGGGACACC ATAAAACAAG GGGGTAACTA CTTATGATCA 1380
1381 CCCCCCCTCG TGCCGAATTC                                           1400
          |    10      |    20      |    30      |    40      |    50      |    60
```

Figure. 13 Composite of the sequence from the 3'RACE derived clone and the Stage 36 cDNA clone.

The two possible intiator methionines are shown in bold but translation is illustrated from the second (reasoning in text). The homeodomain is shown in bold and the 5' and 3' homology boxes are underlined. The arrowhead indicates the intron location.

```
GAATTCACGAGAGCGGCTCAGAGCTCAGCCTGTGGCTCCTGCACTCTGTCTGCCCTGAAGGCTGTCACTGCCATTGCCTGTGCTGGGGCT      90

GCTTACTTCTTGCTACACTGAATACTGGCCCTGCAGAAGGGACACTACCTGATCTGTTAAACTTGCAGGTTGATATTCCTACAGCTTGCC     180

AGTCACATAGACAGTCAGCCCTCGTGCCAACCCGACTTGCTGCCAATCATGCAGTGTGCCTGACCCACAAGCTGTCACCCCAGGGAGTGG     270

GCTACAAGCAGAGAAGACTGTGGCTTTTCCTTAATGAGACCTGTGGATCAGTAGATTTGCAAGATACACTGGTGGTCACACCCTGGTCCC     360

TTTATGTTCAGAGGCCCTCTGGATACCCCCTGGGCAGCAGATGTTTCCCCCAGTGGCCCACGCGGGATCAAGGCTGGACGCTCTGCCCTG     450

                                                                          M  A  A  V
CTAGTGCCCATCTTGGGATCCATTTGGGGAGCTTGGCGCCCAGAGCCAGGAGCCACCCGGTACCGGACACTGGATCCAATGGCAGCTGTG     540

 R  S  S  S  R  L  T  P  F  S  I  Q  A  I  L  N  R  K  E  E  R  A  H  T  F  P  R  L  A  T
CGAAGTTCCAGCAGGCTGACTCCATTCTCTATCCAGGCTATTCTGAACAGGAAAGAGGAGCGTGCCCACACGTTCCCCAGGCTGGCCACT     630

 A  V  S  P  A  C  C  W  R  I  F  G  E  S  E  A  E  G  L  G  S  P  C  G  G  A  P  G  A  G
GCAGTCTCACCCGCCTGTTGCTGGAGGATCTTTGGGGAGAGTGAAGCAGAAGGACTGGGCTCCCCATGTGGAGGGGCCCCTGGGGCAGGA     720

 A  G  G  E  P  S  G  W  D  S  D  S  A  L  S  E  E  G  E  L  G  R  P  G  D  I  G  E  R  K
GCCGGGGGAGAGCCATCCGGGTGGGACTCGGATTCAGCCCTCAGTGAGGAGGGAGAACTGGGAAGGCCGGGGGATATCGGGGAGAGGAAG     810

 K  Q  R  P  L  E  A  R  A  K  G  E  D  E  E  E  T  P  G  C  S  D  C  E  I  G  A  S  V  P
AAGCAGAGGCCGCTGGAGGCCAGAGCCAAGGGGGAGGATGAGGAGGAAACCCCAGGCTGCAGCGACTGCGAGATAGGAGCCAGTGTCCCA     900

 D  P  S  P  P  D  E  D  P  K  C  E  Q  L  M  L  E  P  P  K  Q  R  K  K  R  S  R  A  A  F
GATCCCAGCCCTCCAGATGAGGATCCCAAGTGTGAGCAGCTGATGCTAGAGCCCCCCAAGCAGAGAAAGAAGCGCTCCCGGGCTGCATTC     990
                  ▲
 S  H  A  Q  V  F  E  L  E  R  R  F  N  E  Q  R  Y  L  S  G  P  E  R  A  D  L  A  A  S  L
TCCCATGCCCAGGTCTTTGAGCTGGAGAGACGCTTCAACCACCAGCGATACTTGTCTGGCCCTGAGCGAGCTGACCTGGCTGCCTCCCTT    1080

 K  L  T  E  T  Q  V  K  I  W  F  Q  N  R  R  Y  K  T  K  R  R  Q  M  A  T  D  L  L  A  A
AAACTAACAGAGACCCAGGTGAAGATATGGTTCCAGAACAGGCGCTACAAGACCAAGCGCAGGCAGATGGCTACCGACCTTCTTGCTGCA    1170

 A  P  A  A  K  K  V  A  V  K  V  L  V  R  D  D  Q  R  Q  Y  H  P  G  E  V  L  H  P  S  L
GCCCCTGCAGCAAAGAAAGTTGCTGTGAAAGTGCTGGTGAGGGACGATCAGAGACAGTACCATCCTGGGGAGGTCCTTCACCCTTCACTG    1260

 L  P  L  Q  A  P  Y  F  Y  P  Y  Y  C  A  L  P  G  W  T  L  S  A  A  A  C  S  R  G  T  P
CTTCCCCTTCAGGCCCCTTATTTCTACCCTTATTATTGTGCCCTGCCTGGCTGGACGCTCTCCGCAGCTGCCTGCAGTAGGGGGACACCA    1350

TAAAACAAGGGGGTAACTACTTATGATCACCCCCCCCTTTGCACCCTGGGGGCTTGAATATAGACTCGTTTATTCTGCACCAGCAAAACT    1440

GTGACCAAAAAAGGATTTTGAACAATTGCAGACTTTGTCCATGGTATCCATGTTGTCTATGTGCCGTATAGGTTTTATATCTATATTATA    1530

GCTCCTTATATTGTCTATGGAATGACTAAGCTTTCTGGGCGAGTTGAGGGGTCTGTAAAAAAGGGGATCCCGGGTTCGGACTCTTGGTGG    1620

CCCCCACTGACTGCATGGGGCTGTTTCTAGGTTCCATGGGTGTCCAGGGCTACGAATGTATAGTACAGAGCTTCCTTATCGCTTTACCAT    1710

TCCCCATCCTGCCTCGTTCTTACTCTCCTCATCATCCATTAGATTGACGTCATCTAAAGGGGTCTGCATAGGTAGCAATCCATAGAGATG    1800

TACTACGTTAGGTGGGGAGGTAAGACATTTTTCTAGCACTTTTATTCTCTTGCAACCTGTGGCCTTTCTCTTGCAACTATCGTCTCTGCC    1890

TTTCCCAGACAGTTGGGAATATAAATCCAACTACAAATGGGGAACCTTTGGATGGATAATCTGGCAGAGATACTTGCTAGAGCTACGTGC    1980

TTTTATCTCCGCTCAGCTCAGTGCCTTTCCCGAAAAACACTTAGACACAGGCATCTGCTTCATTGTTGCCTTATTTATAAACACAAGATG    2070

CAATCCTTTAATAATCCCGCAGTTTGTAGATTCACAAGAATCTGGCACTGCGGCATCAGCTGCATTACTGTATGTGCTTTTCCAATGACA    2160

TTGTGATAAGTGTTTTTTAATACTTTGTTTTTTTTCTAATTTATAAAGTATAGTTTAACTAAAAAAAAAAAAA                    2233
```

to probe a southern blot of the five genomic positives each digested with Bam H1, Pst 1 and Xba 1 in order to identify which positives contained regions which could hybridise to both. One such positive was identified and was digested with Xba 1 generating two fragments, one of approximately 5Kb and the other approximately 3Kb, which were subcloned into puc 18. A restriction enzyme map of these subclones was generated by digesting each of them with Bam H1, Pst 1, Eco R1, Bgl 1, Sma 1, and Sty 1 as single digests and in combination with Xba 1. Duplicate southern blots of these digests were probed with either the Eco R1/Bam H1 probe or the Sma 1/Eco R1 probe. This map suggested the presence of an intron upstream of the homeobox and this was confirmed by sequencing across the exon/intron junction in appropriate M13 subclones. The intron/exon junction is indicated in figure 13.

In an attempt to isolate upstream regulatory elements which may control the expression of the clone a southern blot of all of the genomic clones was screened with a probe corresponding to the 3' untranslated region of the cDNA clone; unfortunately none of the clones were positive in this region. A second method for the isolation of upstream sequence, inverse PCR, was used by Paul Kriegs group and produced a clone with an 200bp insert. This clone was kindly sent to us and following appropriate subcloning was fully sequenced. Close analysis of the sequence failed to reveal any promoter elements (TATA box or GC rich regions). The inverse PCR method could be used to gain further upstream sequence but unfortunately there are very limited restriction enzyme sites in the sequence, which can be exploited for this method. Another disadvantage of the inverse PCR technique is that the length of novel sequence isolated is highly variable and unpredictable. The screening of a genomic library with a probe derived from the 5'untranslated region of the positive clone would, therefore, be the preferred method in any attempt to isolate further 5'UT sequence.

3.2.2 Database analysis and sequence comparisons

When the cDNA clone was first isolated a BLAST database search with the sequence produced the best match over the homeodomain region with *Drosophila Bagpipe* (76%

identity), followed by X*Nkx-2.5* (73% identity). However, since the initial cloning a gene named *Xenopus Bagpipe* (*XBap*) was published by Newman *et al*, 1997 which is identical in sequence to our cDNA clone. *NK3* relatives have now also been identified in mouse, human and *planaria*.

Outside of the homeodomain the amino acid sequences of *NK* class homeobox genes are highly divergent except for two additional boxes of homology one located upstream of the homeobox (TN domain) and a second, comprising a conserved decapeptide (NK2 box), downstream. These homologous regions do not lie at a defined distance from the homeobox in the different genes, but they are easily identified and can be aligned by looping out intermediate nucleotides from some sequences. Figure 14 is an alignment of the three conserved regions of all the NK3 related proteins identified to date.

Murine Bapx1 is identical to XBap across the homeodomain and additional regions of homology at the nucleotide level identified using the GCG programs compare and dotplot suggest that murine *Bapx1* is an ortholog of *XBap* (Genetics Computer Group, 1994). Figure 15 illustrates compare/dotplots of *XBap* in comparison with *Drosophila bagpipe*, murine *Bapx1* and murine *Nkx-3.1*. Homology at the nucleotide level between *Drosophila bagpipe* and *XBap* is limited to the homeobox with no apparent homology in the region of the TN or NK2 domains. Nkx-3.1 also shows no homology in the region of the TN domain but the homology across the homeobox extends C-terminally to include the NK2 box.

The nucleotide sequence of the homeoboxes of *Drosophila NK2-4* genes and several vertebrate NK class genes were used to generate a tree (Fig.16) using the GCG program clustal, which utilises the neighbour joining method to compare the nucleotide sequences of the homeodomains, and tree-draw to graphically represent the results (Genetics Computer Group, 1994). This is not a true phylogenetic tree with a root and a time scale as it was produced by a series of pairwise alignments between all the sequences in which the route producing the least number of branches is followed until only two sequences

Figure 14. Comparison of the *Xenopus* Bagpipe 'TN' domain, homeodomain and 'NK2' domain sequences with NK3 class proteins identified in other species.

Gene names are prefixed with a species abbreviation: m, mouse; Dro, *Drosophila*; pn, *planaria*; p, *pleurodeles*; ch, chick; X, *Xenopus*; h, human; d.r. zebrafish *danio rerio*; r, rat.

Murine Bapx1 and human Bapx1, Triboli et al, 1997, murine Nkx-3.1, Bieberich et al, 1996, human Nkx-3.1, Prescott et al, 1998, Zampogna, Newman et al, 1999, *planaria* Bagpipe, Balavoine et al, 1996, *pleurodeles* Nkx-3.2 and Nkx-3.3, Nicolas et al, 1999, NK-3, Kim et al, 1989.

```
          XBap   TPFSIQAILNR
      XZapogna   .S....D..A.
  Mus Nkx-3.1    .S.L..D..RD
    Mus Bapx1    ..........K
  Dro bagpipe    .....ND..T.
   P  Nkx-3.2    ..........K
```

```
          XBap   KKRSRAAFSHAQVFELERRFNHQRYLSGPERADLAASLKLTETQVKIWFQNRRYKTKRRQ
      XZapogna   ............Y......SL.....................................KL
    Mus Bapx1    ..................................................
  Mus Nkx-3.1    Q.........T..I....K.S..K...A....H..KN.....................K.
       Hum Bap   ..................................................
      XZapogna   ............Y......SL.....................................KL
       Pn Bap    ...T......T..Y.....G.......S...E..R..R.S...I............K..
    P  Nkx-3.2   ..................................................
    P  Nkx-3.3   ............Y......SL.........A.....................K.
  Dro bagpipe    ..................A.........SEM.K..R.....................K.
```

```
          XBap   PAAKKVAVKVLVRDD
      XZapogna   VPTR....R...K..
    Mus Bapx1    ..............
    P  Nkx-3.2   G.............
    P. Nkx-3.3   S..RR..IR......
  Dro bagpipe    G.S.R.PIQ....E.
```

Figure 15. Compare/Dotplots of *XBap* against *Drosophila bagpipe*, murine *bapx1* and murine *Nkx-3.1*

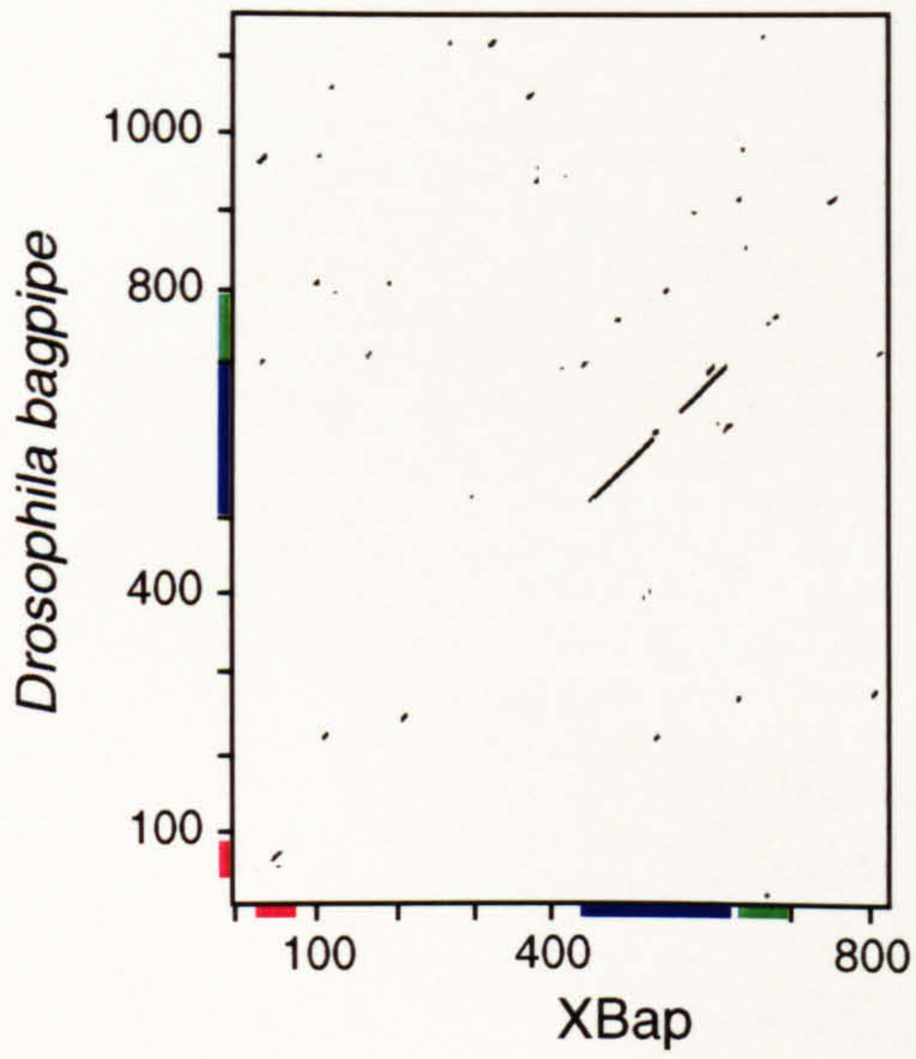
The GCG programs compare and dotplot were used to detect regions of homology within the open reading frame, at the nucleotide level, between *XBap*, *Drosophila bagpipe* (Kimand Nirenberg,1989), Murine *bapx1* (Triboli *et al*, 1997) and murine *Nkx-3.1* (Bieberich *et al*, 1996).

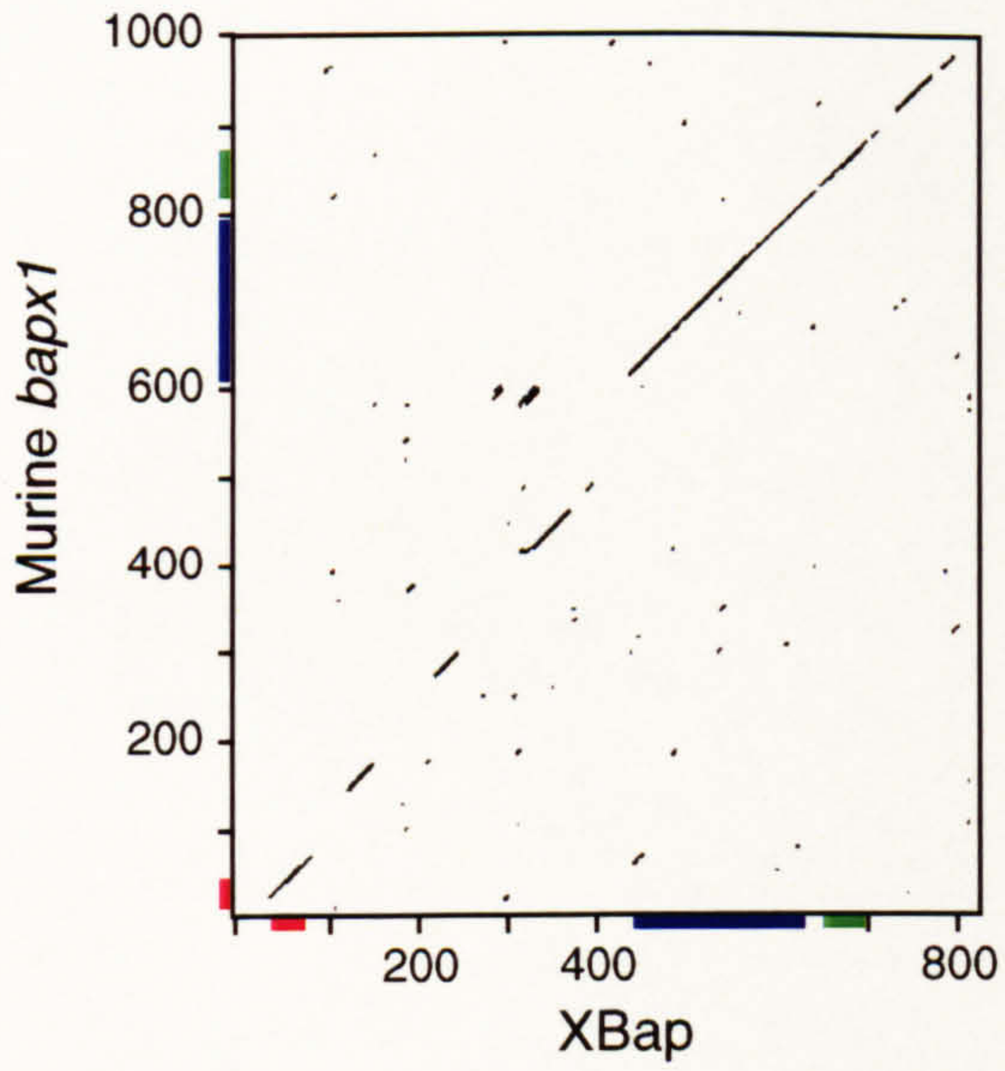
Key;

Red: TN domain

Blue: Homeodomain

Green: NK2 box


Drosophila bagpipe
1000
800
400
100
100
400
800
XBap


Murine bapx1
1000
800
600
400
200
200
400
800
XBap

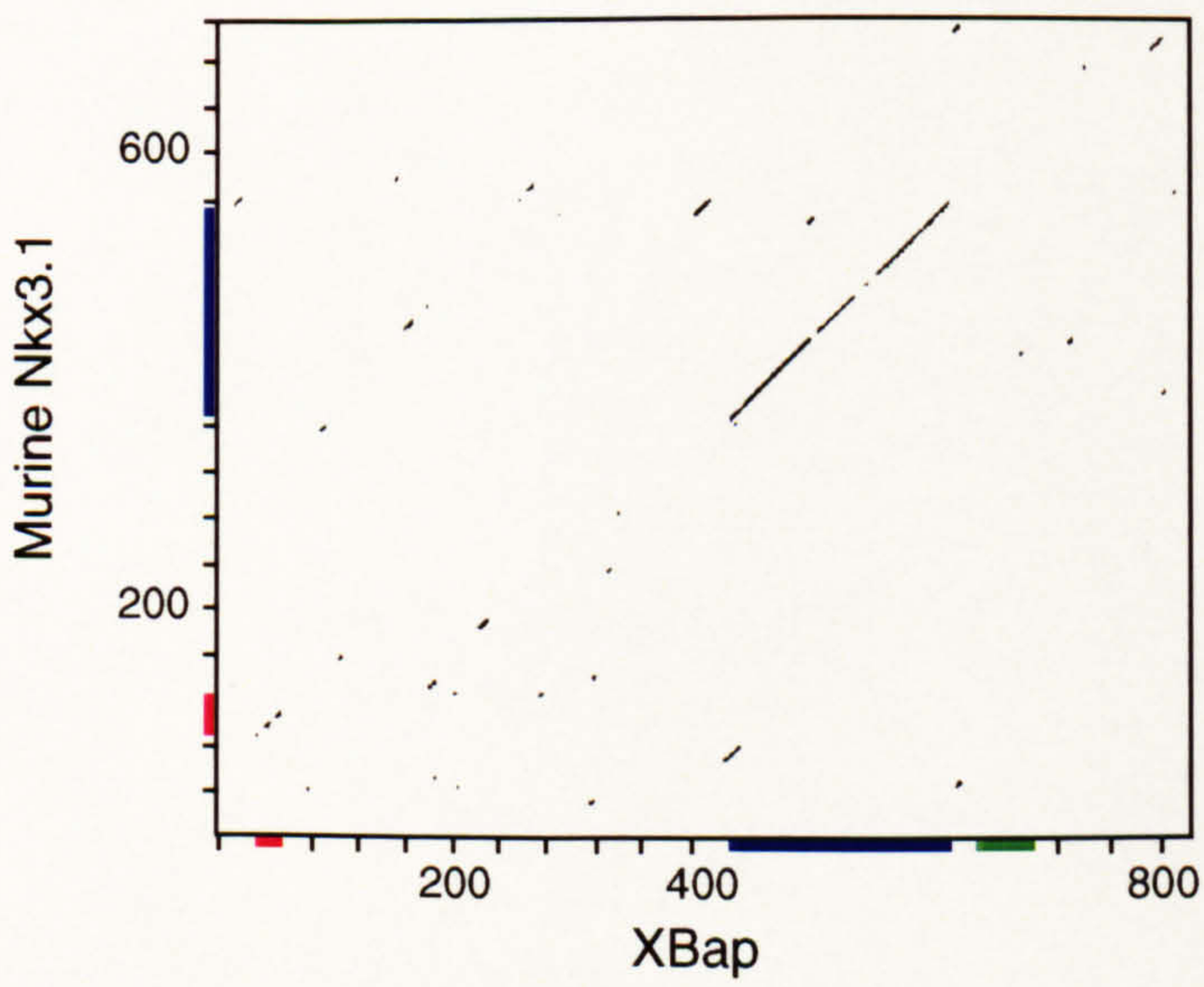
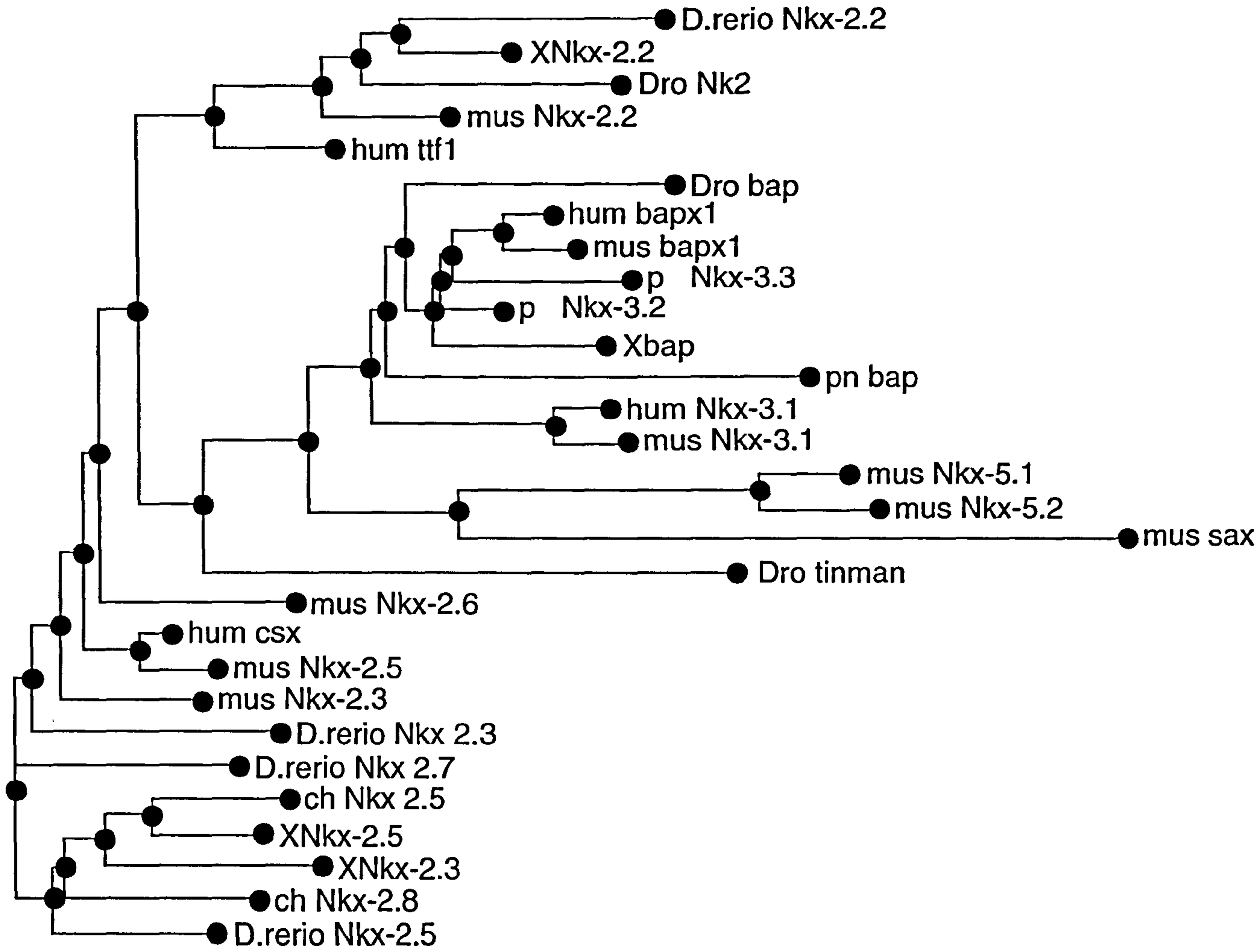

Murine Nkx3.1
600
200
200
400
800
XBap

Figure. 16 Tree diagram of NK-2, NK-3 and NK-4 class genes from invertebrates and vertebrate species.

This tree as generated using clustall program which utilises the neighbour joining method to compare the nucleotide sequences of the homeodomains, and treedraw to graphically represent the results.

Gene names are prefixed with a species abbreviation; D. rerio, *Danio rerio* zebrafish; X, *Xenopus*; Dro, *Drosophila*; mus, murine; hum, human; pn, *planaria*; p, *pleurodeles*; ch, chick.

Sequence references: Murine Nkx-2.5, Lints et al, 1993, human csx, Turbay et al, 1996, chick Nkx-2.5, Schultheiss et al, 1995, *Xenopus* Nkx-2.5, Tonissen et al, 1994, D.rerio nkx-2.3, nkx-2.5 and nkx-2.7, Lee et al, 1996, *Xenopus* Nkx-2.3, 1995, chick Nkx-2.8, Brand et al, 1995, murine Nkx-2.6, Lints et al, 1996, NK-1, NK-2, NK-3 and NK-4, Kim et al, 1989, murine Nkx-2.2, Hartigan et al, 1996, *Xenopus* Nkx-2.2, Hartigan et al, 1996, murine bapx1, Triboli et al, 1997, murine Nkx-3.1, Bieberich et al, 1996, human bapx1, Yoshiura et al, 1997, *planaria* bagpipe, Balavoine et al, 1996, *pleurodeles* Nkx-3.2 and Nkx-3.3, Nicolas et al, 1999, murine Sax1, Schubert et al, 1995, murine Nkx-5.1 and Nkx-5.2, Bober et al, 1995, lox 10, Nardeli-Haefiger et al, 1993, rat ttf-1, Guazzi et al, 1990,human ttf-1, Hamdan et al, 1998, *planaria* dth-1 and dth-2, Garcia-Fernandez et al, 1991, human Nkx-6.1, Inoue et al, 1997.


D.rerio Nkx-2.2
XNkx-2.2
Dro Nk2
mus Nkx-2.2
hum ttf1
Dro bap
hum bapx1
mus bapx1
p Nkx-3.3
p Nkx-3.2
Xbap
pn bap
hum Nkx-3.1
mus Nkx-3.1
mus Nkx-5.1
mus Nkx-5.2
mus sax
Dro tinman
mus Nkx-2.6
hum csx
mus Nkx-2.5
mus Nkx-2.3
D.rerio Nkx 2.3
D.rerio Nkx 2.7
ch Nkx 2.5
XNkx-2.5
XNkx-2.3
ch Nkx-2.8
D.rerio Nkx-2.5

remain. The tree places the *Drosophila NK2* and vertebrate *Nkx-2.2* genes in a cluster, which correlates well with their shared patterns of expression in lung and brain (Nirenberg *et al*, 1995). The vertebrate *Nkx-2.3* to *2.7* genes are in a distinct cluster and the branch containing them is closer to the *Drosophila NK4* gene (*tinman*), whose functions they appear to emulate, than to their namesake gene *NK2* (Harvey, 1996). The *NK3* class genes are all clustered together with *XBap* among them. This type of analysis demonstrates that the current nomenclature of this family is not functioning to represent the true relationships between the genes. It should be replaced by a system that recognises that there is not a clear division of the large number of vertebrate genes into four groups to correlate with each of the *Drosophila NK1-4* genes.

Considering the presence of more than one NK3 gene in mouse, humans and *planaria*, as well as the family of genes that appear to represent *tinman* in all vertebrates to date, it seems likely that *Xenopus* will prove to have more than one *NK3* class gene. Although only one *bagpipe* relative was identified in this screen it may be that at stage 36 of *Xenopus* development only one type of *NK3* gene is expressed. As *XBap* is expressed relatively late in development, it is not detected until stage 19 it is possible that another *NK3* gene may be expressed at an earlier stage. Although a screen of a stage 17 library failed to identify any positives each of the *NK3* class genes studied to date have shown low levels of expression in very small and restricted areas of tissue (Newman *et al*, 1997, Triboli *et al*, 1997, Sciavolino *et al*, 1997 and He *et al*, 1997). The alignment of all *NK* class genes (Fig. 7) shows that the *NK3* class genes are most distinguishable from their *NK2* class relatives by the sequence of the N-terminal arm. A 3'RACE screen using primers designed on the basis of this region followed by a cDNA library screen may successfully isolate additional *NK3* related genes.

3.2.3 Translation initiation site

The *XBap* sequence reveals one open reading frame containing the homeodomain but there are two possible initiator methionines. Neither of these methionines fit well to the Kozak consensus sequence for initiation sites in higher eukaryotes (CCa/gCCAUGG,

Kozak, 1987). More recently Cavener and Ray (1991) have analyzed sequences flanking the translation initiation sites from a number of eukaryotic cDNAs. This has revealed a few key similarities amongst vertebrate initiation sites. Of particular note are:

i) At positions +4 and +5, G and C are preferred respectively, the upstream methionine in the XBap sequence has a T at both of these positions but the downstream Met has G and C at +4 and +5.

ii) At the -2, -1 and +4 positions the most frequent nucleotides are CC···G. Again only the downstream methionine conforms to this observation.

iii) Triplet frequencies at the -3,-2,-1 position has been analysed. The UUU of the upstream methionine is observed in 0.1% of cases whereas the CCA of the downstream methionine is observed in 0.8% of cases.

Although neither of these methionines have surrounding sequences thought to produce optimal initiator sites, (for example the C at -3 of the downstream methionine is unusual), these observations would seem to suggest that the downstream methionine is most likely to be the *in vivo* initiator.

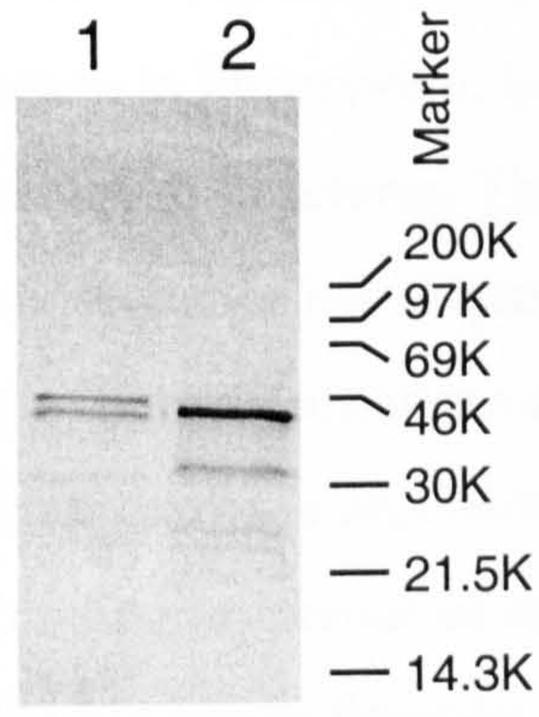
in vitro translation of synthetic RNA produced from the Orig/pT7 clone, which contains both methionines, indicated that each of the methionines are capable of translation initiation *in vitro* (Fig.17). However, in all of the vertebrate NK members studied to date, the initiator methionine is just upstream of the amino-terminal conserved box. This, in addition to the sequence analysis, strongly suggested that the downstream methionine is the most likely *in vivo* initiator of this protein. It was, therefore, this methionine that was used to produce a clone corresponding to the open reading frame of *XBap* (ORF/Myc/pβ).

3.3 Future work

This chapter has described the cloning of *Xenopus Bagpipe*, the first *Xenopus* member of the *NK3* class of homeobox genes. One of the first steps following the cloning of a novel gene is the determination of its expression pattern throughout embryogenesis. Preliminary

Figure 17. *In vitro* translation proteins produced from two *XBap* constructs.

1. *In vitro* translation product from the Orig/pT7 construct which contains two potential initiator methionine codons.
2. *In vitro* translation product from the ORF/Myc/pT7 construct from which the 5' ATG codon has been eliminated.


1
2
Marker
200K
97K
69K
46K
30K
21.5K
14.3K

experiments aimed at the determination of the expression pattern by RNase protection and RT-PCR had commenced. However, Newman *et al*, who had cloned the *XBap* gene simultaneously by independent means, published a detailed analysis of the embryonic expression pattern of *XBap* (as described in the introduction) whilst this project was ongoing (Newman *et al* ,1997).

Xenopus Bagpipe is expressed in the posterior foregut, which approximately corresponds to a domain of *bagpipe* expression in *Drosophila*, however, *XBap* is also expressed at high levels in developing craniofacial structures. This difference in expression pattern may suggest that the homeobox sequence has acquired a novel developmental function during evolution. The binding of homeoproteins to specific DNA binding sites in the regulatory regions of target genes can have an effect on the transcription of that gene, characterisation of the DNA binding properties of each homeoprotein is therefore a good first step in elucidating possible biological function. It was decided that the most fruitful direction for further study would be the determination of the DNA binding site of *XBap* and characterisation of its DNA binding characteristics.

Chapter 4

Selection of the XBap DNA binding site

4.0 Chapter 4: Selection of the XBap DNA binding site

4.1 Introduction

Previous studies have revealed a surprising lack of variability in the DNA binding sites of homeoproteins; the majority recognising DNA sites with a 5'-TAAT-3' core with variation occurring in the base 5' upstream and the base pair 3' downstream (5'-NTAATNN-3'). NK class proteins studied to date, however, bind specifically to sites containing a 5'-CAAG-3' core (Damante *et al*, 1994, Chen *et al*, 1995). Support for such a site being bound specifically by NK2 class proteins *in vivo* is provided by the fact that Nkx-2.1 was identified on the basis of its interaction with 5'-CAAG-3' sites in the thyroglobulin promoter (Guazzi *et al*, 1990). It was, therefore, of interest to identify whether XBap has a binding site similar to the NK2 class proteins, the more common 5'-TAAT-3' motif, or a distinct motif.

In vitro binding site selection systems are based on a polymerase chain reaction with variable input DNA and methods of purification of the protein-DNA complexes. The selection procedure may be 'weighted' if there is already information about the proteins DNA binding i.e. the pool of oligonucleotides may all contain a conserved core outside of which the sequence is random. If there is no information about possible binding motifs of the protein then a pool of random oligonucleotides is used as the starting material. Either an EMSA assay or immunoprecipitation by specific antibodies or via a tag can be employed for purification of protein/DNA complex.

In order to define the XBap DNA binding site it was decided to use the DNA binding site selection system developed by Pollock and Treisman, (1991 B). Protein extract, or *in vitro* translated proteins, are first incubated with a pool of random sequence oligonucleotides. The protein/DNA complexes are purified by immunoprecipitation, and the DNA amplified by PCR. This DNA is then used in further rounds of binding, immunoprecipitation, and amplification, until specific binding is detectable in an EMSA assay. In Pollock and Treisman's study, SRF from transfected cell extracts, or *in vitro* translated SRF, was observed to select authentic SRF sites over binding sites recognised

by other transcription factors in the upstream regulatory regions of SRF target genes thus providing evidence that sites identified by this *in vitro* site selection procedure can be a true representation of the sites bound by a protein *in vivo* (Pollock and Triesman, 1991 B).

The study of the *in vitro* DNA binding preferences of PPARα/RXRα heterodimers is just one example among many, in which *in vitro* studies have been shown to be representative of *in vivo* binding. The oligonucleotides selected all contained two or three copies of the half-site sequence AGGTCA. These sites were most often organised as a direct repeat with one nucleotide spacing, the established configuration of regulatory elements in peroxisome proliferator target genes (Castelein *et al*, 1997). A similar method, which used EMSA to purify protein/DNA complexes instead of immunoprecipitation, was used to study the DNA binding specificity of the CCAAT-binding factor CBF/NF-Y. This analysis revealed that CBF binding to DNA requires not just the CCAAT sequences but also additional sequences immediately flanking both ends of the motif. Most of the CCAAT motifs present in various higher eukaryotic promoters were found to correspond to the extended CBF binding sites that were selected *in vitro* (Bi *et al*, 1997). Of greatest relevance to this study is the fact that the two Tinman binding sites identified in the enhancer of D-mef2, which have been shown to be essential for D-mef2 expression, match the *in vitro* consensus determined for Tinman (Gajewski *et al*, 1997).

4.2 Results

4.2.1 Immunoprecipitation

In order to use the site selection procedure it was necessary to tag XBap with an epitope of a commercially available antibody so that it could be immunoprecipitated. The full open reading frame of *XBap* was subcloned into the pβTAG vector (see Materials and Methods) in order to introduce a myc tag, which is recognised by 9E10 antibody, at the amino terminus (ORF/myc/pβ) . Tagged protein was produced by *in vitro* translation of synthetic RNA in reticulocyte lysate. An immunoprecipitation reaction was performed under the conditions used by Pollock and Treisman (1991). The tagged XBap protein is

immunoprecipitated with 9E10 monoclonal antibody (Figure 18, Lane F). However, XBap protein also adhered non-specifically to protein-A sepharose beads in the absence of antibody. In an attempt to improve the specificity of immunoprecipitation, a range of increasingly stringent wash conditions were tested, (Fig.18). Protein-G-sepharose beads pre-cleared with BSA were used in place of protein-A as 9E10 antibody can bind directly to protein-G without the need for anti-rabbit IgG (XBap may be interacting with the beads via the anti-rabbit IgG). After binding, Dignams buffer D + 0.1% NP40 washes were replaced with more stringent WEEB and RIPA buffers (total of four washes) (Lanes C-F) and one set had an additional distilled water wash (Lanes G and H). All reactions were carried out in silicanised tubes and the reactions were transferred to new tubes just before elution of the protein. Despite all these measures Lanes E-H illustrate that as the stringency is increased, both precipitation with, and without antibody, decreases to the same degree. This phenomenon has also been observed in the laboratory with XNkx-2.5 and XNkx-2.3 (Mohun *et al*, unpublished data). In summary, these experiments indicated that the construct (ORF/myc/pβ) was suitable for use in binding site selection, as the primary concern is to immunoprecipitate XBap protein associated with DNA, but not for co-immunoprecipitation experiments for which it is imperative that there is no non-specific binding of the proteins.

4.2.2 Binding site selection

Binding site selection was carried out using the method as described by Pollock and Treisman, (1991 B), using *in vitro* translated XBap. Firstly the oligonucleotide R76, which contains 26bp of random sequence, was used in a binding reaction with XBap protein followed by immunoprecipitation of the protein/DNA complexes. The random oligos for site selection have BamH1 at one end and EcoR1 at the other.

. The initial input of DNA is 0.4ng, which represents approximately 5 X 10^9 molecules. The bound oligonucleotides eluted from the protein were amplified by PCR and used in the next round of binding, immunoprecipitation and amplification. After four rounds of selection the radiolabelled PCR products of each round were used in an EMSA assay with XBap protein (Fig. 19). A

Figure 18. Immunopreciptation of XBap

Immunopreciptation of myc-tagged XBap, produced by *in vitro* translation from synthetic RNA, under a range of conditions. Protein G Sepharose was used in all of the reactions.

A 5µl XBap reticulocyte lysate product

B 5µl H_20 reticulocyte lysate product

C H_20 reticulocyte lysate, no antibody control, alternate WEEB and RIPA washes

D H_20 reticulocyte lysate + 9E10, alternate WEEB and RIPA washes

E XBap, no antibody, alternate WEEB and RIPA washes

F XBap + 9E10, alternate WEEB and RIPA washes

G XBap, no antibody, alternate WEEB and RIPA washes + one H_2O wash

H XBap + 9E10, alternate WEEB and RIPA washes + one H_2O wash

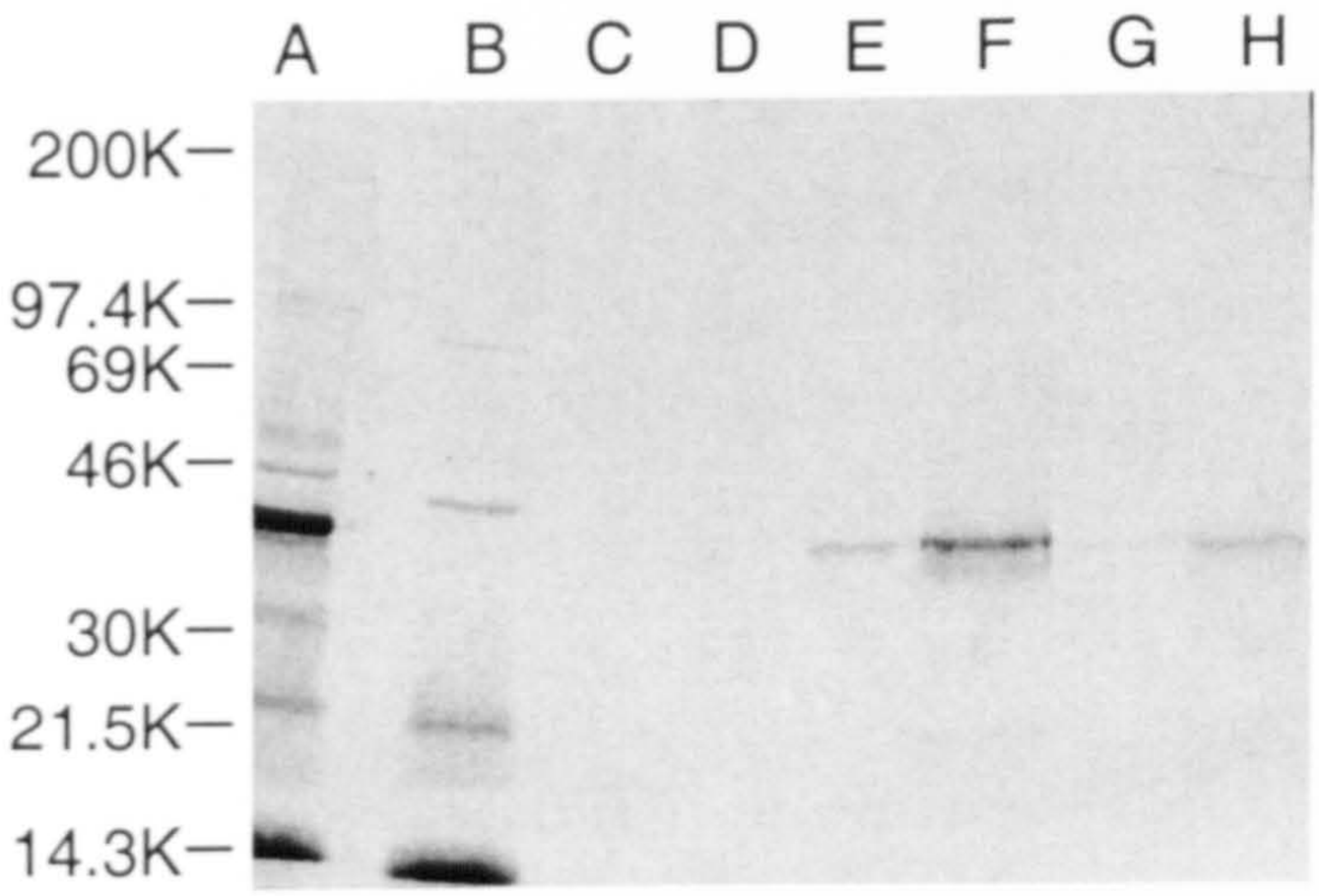
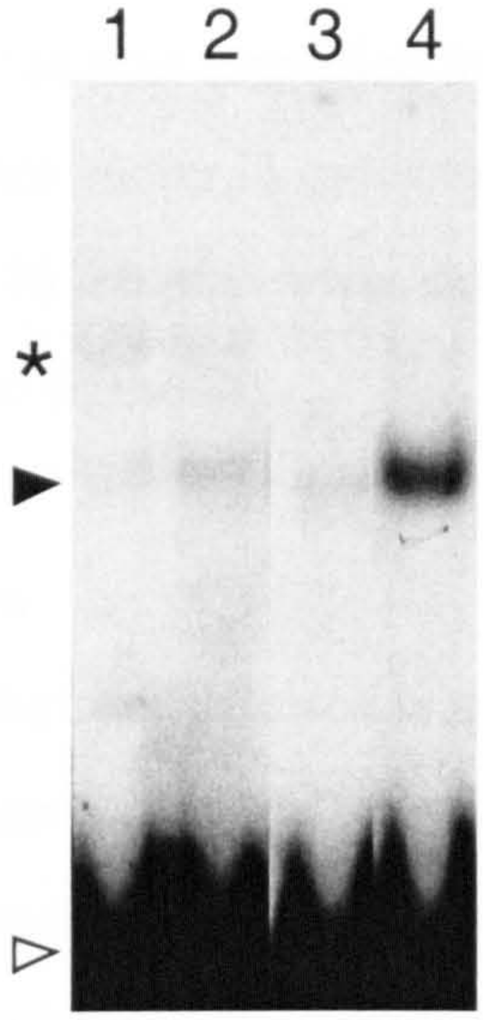

A
B
C
D
E
F
G
H
200K—
97.4K—
69K—
46K—
30K—
21.5K—
14.3K—

Figure 19. Bandshift with XBap protein and the products of each round of binding site selection.

Full length XBap produced by *in vitro* translation of synthetic RNA was used in each binding reaction. Lanes 1, 2, 3, and 4 illustrate the amount of specific complex formed with probes from the first four rounds of selection.



► = specific complex

▷ = free probe

* = background

complex was formed with the 4th round probe indicating that some selection for optimal sites had occurred but to obtain even further enrichment, this complex was excised from the gel, the DNA extracted and used for a further 3 rounds of selection. A second bandshift carried out using probe from rounds 5, 6 and 7 indicated that the extra three rounds had significantly enriched for DNA able to form a complex with XBap. DNA from the seventh round of selection was amplified by PCR, digested with BamH1 and EcoR1 then subcloned into an M13 mp18 Bam vector in order to force dimers and reduce the amount of sequencing. The sequence of 46 templates was determined and this yielded 97 potential sites, which are listed in figure 20.

4.2.3 Analysis of selected sites

The most notable feature of all the sites selected is that they contain one very highly conserved motif and a second more degenerate version of the motif. The spacing and orientation of the two motifs is variable. In figure 20, however, the sequences are listed in order of spacing/orientation occurrence which demonstrates that the most common arrangement of the two motifs is a 5'-TAAG-3' orientation with a three base pair separation (71% in a positive orientation and 32% with 3 nucleotide spacing). The 5'-TAAG-3' core is almost invariant in both motifs (99% 1st site, 91% 2nd site), and 5'-TTAAGTGG-3' is the most common sequence of the first site (59%). In both the first and the second motif, however, it is the first base and last trio that are most subject to variation. Analysis of the sites revealed that only a limited number of trios occur and that in some cases a trio has a preference for a particular first base. Figure 21 A, is a table constructed to illustrate the occurrence of first base and final trios in the selected sites. Figure 21 B, indicates the three most common sites actually selected.

As 59% of the most conserved motifs selected are of the 5'-TTAAGTGG-3' sequence, and this is also a common sequence among the second sites, with a three base pair gap and a positive orientation as the predominant arrangement of the two sites, 5'-TTAAGTGG---TTAAGTGG-3' has been designated as the 'T2' site. An oligonucleotide of this sequence was used in a preliminary EMSA assay to demonstrate that the selected

Figure 20. Sequence of potential binding sites in order of most common spacing and orientation.

Each site contains two copies of a core motif. The first, most conserved motif of each potential site, is highlighted with a blue background. The second, more degenerate motif, is highlighted with a green background if it is in a positive orientation relative to the first, or red, if it is in a negative orientation.

GGATCCTGTCGCTGTTATGAAGTTAAGTGGTTGTTAAGAGGCGAATTC	3+
GAATTCGCCTCCTCCTAAGCACCCCTTAAGTGCTTTTCGACAggatcc	3+
GGATCCTGTCGCATCGTAAGTGGTAGATAAGTGGTCGGAGCggaattc	3+
GGATCCTGTCGCATACTTAAGTGGCCATAAAGTGTGCGAGGCggaattc	3+
GAATTCGCCTCGATTTAAGTGGTCCTTAAGTGCTCCCCGACAggatcc	3+
GGATCCTGTCGGGGGGATATGTTAAGTGGCTCTTAAGGAGGCggaattc	3+
GGATCCTGTCGTTAAGTAGCTCTTAAGTGTTCCATGAGAGGCGAATTC	3+
GAATTCGCCTCGATTTAAGTGGTCCTTAAGTGCTCCCCGACAGGATCC	3+
GGATCCTGTCGAGCCTTAAGTACTTACTAAGTGGTTGGAGGCGAATTC	3+
GGATCCTGTCGTTAAGTAGCTCTTAAGTGTTCCATGAGAGGCGAATTC	3+
ggatccTGTCGTTAAGTGGGTTGTAAGTGGATAAAATGAGGCGAATTC	3+
ggatccTGTCGTTAAGTGGGTTGTAAGTGGATAAATGAGGCGAATTC	3+
GGATCCTGTCGCTGTTATGAAGTTAAGTGGTTGTTAAGAGGCGAATTC	3+
ggatccTGTCGAGCCTTAAGTACTTACTAAGTGGTTGGAGGCGAATTC	3+
GGATCCTGTCGGTAAGTGGACGTTAAGTGGTTTAGGTGAGGCGAATTC	3+
GGATCCTGTCGAAATAAGAGGCACTTAAGTACCTTGTGAGGCGAATTC	3+
ggatccTGTCGGTAAGTGGACGTTAAGTGGTTTAGGTGAGGCGAATTC	3+
ggatccTGTCGGCCATTAAGTGGTACTTAAGTGGGACGAGGCGAATTC	3+
ggatccTGTCGGTAAGTGGACGTTAAGTGGTTTAGGTGAGGCGAATTC	3+
GGATCCTGTCGATAGGTTAACTGGTATTTAAGCGGTGGAGGCGAATTC	3+
GGATCCTGTCGGTAAGTGGGTCGTAAGTGGCTAGAACGAGGCGAATTC	3+
gaattccGCCTCGATTTGAAGTGGCCCTTAAGTGCTCCCCGACAGGATCC	3+
GGATCCTGTCGTTTAAGTGGGTGTTAAGTACTTAGAGGAGGCgaattc	3+
GGATCCTGTCGACTTAAGTGGTATTGAAGTGGACGAGCGAGGCggaattc	3+
GGATCCTGTCGAGATAAGTGGATTTTAAGTGGGGTGTGAGGCggaattc	3+
GGATCCTGTCGAGCCTTAAGTACTTACTAAGTGGTTGGAGGCggaattc	3+
GAATTCGCCTCATAAGTGTTTCTTAAGTGCCTTCTCCGACAGGATCC	3+
GGATCCTGTCGTTTAAGTGGGTGTTAAGTACTTAGAGGAGGCgaattc	3+
GGATCCTGTCGGTAAGTGGACGTTAAGTGGTTTAGGTGAGGCggaattc	3+
GGATCCTGTCGTTAAGTGGGTTGTAAGTGGATAAAATGAGGCGAATTC	3+
GAATTCGCCTCTGGAGTAAGTACCCTTAAGTGGACTACGACAGGATCC	3+
GGATCCTGTCGAGCATCTTTAAGTGGTTAAGTGGTTTGAGGCGAATTC	0+
GGATCCTGTCGTTGATAAGCGGTTTTTAAGTGGTTAAGAGGCggaattc	0+
GAATTCGCCTCGCCTTAAGTAGTTAAGTGGACGGCATCGACAGGATCC	0+
GGATCCTGTCGAGGTTAAGTACTTAAGTGGTCGTCCTGAGGCGAATTC	0+
GGATCCTGTCGGGGACTAAGTGGTTAAGTGGTAGTCTGAGGCggaattc	0+
GGATCCTGTCGTCCGTTAAGTGGTTAAGTGCAAGCAAGAGGCGAATTC	0+
GGATCCTGTCGTCCGTTAAGTGGTTAAGTGCAAGCAAGAGGCGAATTC	0+
ggatccTGTCGAGCATCTTTAAGTGGTTAAGTGGTTTGAGGCGAATTC	0+
ggatccTGTCGATAGACACTTAAGTATCCTAAGTGGTGAGGCGAATTC	5-
GGATCCTGTCGACTTAATGGTATTAAGTGGACGAGCGAGGCggaattc	5-
GGATCCTGTCGTTAAGTGGGTATAGTACTTATGGCTTGAGGCggaattc	5-
GAATTCGCCTCAGAGCACTTAGCGACGTTAAGTGGCTCGACAGGATCC	5-
GAATTCGCCTCCTAAGGGCTTTAAGTACTTGCTGTGGCGACAggatcc	5-
gaattccGCCTCTCACTTCCCCCCTTAAGTGGTTGTGACGACAGGATCC	5-
gaattccGCCTCTTAAGTGGTCACTTAAGTACTTATACCGACAGGATCC	5-
ggatccTGTCGGAGTTAAGTAGGCACTTAAGTGGCCAGAGGCGAATTC	4+
GAATTCGCCTCTTAAGTGGTCACTTAAGTACTTATACCGACAGGATCC	4+
GGATCCTGTCGCTATTAAGTGCAATACTAAGTGGATGAGGCggaattc	4+
GAATTCGCCTCTAAGTGCCTATTTAAGTGCCTGAATACGACAggatcc	4+
gaattccGCCTCTAAGTACCTCGCTAAGTGGCCTTTGTCGACAGGATCC	4+
gaattccGCCTCTAAGTACCTCGCTAAGTGCCCTTTGTCGACAGGATCC	4+
GGATCCTGTCGGTCTGTAAATTAAGTGGGTACTTAAAGAGGCGAATTC	0-
GGATCCTGTCGGACCGATAAGTGGGCACTTAGCGTGGGAGGCGAATTC	0-
GAATTCGCCTCAAAATGTTTAGTTAAGTGGGTACTTACGACAggatcc	0-
gaattccGCCTCCCTTGTCTGCATAAGTGGTCACTTAGCGACAGGATCC	0-
gaattccGCCTCACTTAACGACTTTAAGTGCCCATCAGCGACAGGATCC	0-
GAATTCGCCTCGATTACTAAGTGGTACTTAATGTCCCCGACAggatcc	-1-
ggatccTGTCGAGCCTTAAGTACTTACTAAGTGGTTGGAGGCGAATTC	-1-
GGATCCTGTCGCTTCTTATCGGTAAGTGGTACTTATACAGGCGAATTC	-1-
GGATCCTGTCGGTTAAGTGGTACTTAAGCAAACGCCAGAGGCggaattc	-1-
GGATCCTGTCGTAGCACTTAAAGACTTAAGTGGGCAGGAGGCGAATTC	4-
ggatccTGTCGTAGCACTTAAAGACTTAAGTGGGCAGGAGGCGAATTC	4-
GGATCCTGTCGCAAGAGTACTTAGTGCACTAAGTGGGGAGGCGAATTC	4-
ggatccTGTCGCTGCAAAAGGCTTAAGTGGTCTTAAGAGGCGAATTC	2+
ggatccTGTCGGAGTTAAGTGGCTTTAAGTATGTTGCGAGGCGAATTC	2+
GGATCCTGTCGACTTAAGTGGTATTAAGTGGACGAGCGAGGCggaattc	2+
GGATCCTGTCGAAAAACTAAGTGGACTAAGTGGTTAGGAGGCGAATTC	1+
ggatccTGTCGCCATGCTTAAGTACCTTAAGTGGTCAGAGGCGAATTC	1+
GGATCCTGTCGCCATGCTTAAGTACCTTAAGTGGTCAGAGGCGAATTC	1+
ggatccTGTCGTCCTTAAGTGGACCACTTAGTATTTTGAGGCGAATTC	1-
GAATTCGCCTCGAGAGCTCTCTTAATATAAGTGTTTCCGACAggatcc	1-
GGATCCTGTCGTCCTTAAGTGGACCACTTAGTATTTTGAGGCGAATTC	1-
GAATTCGCCTCGATTTAATGGTCCTTAAGTGCTCCCCGACAggatcc	-4-
gaattccGCCTCTTAAGTGGCTACTCTAAAGCACTTTTCGACAGGATCC	-4-
GGATCCTGTCGGACAGTATATTAAGTGGCTTTTGAGTGAGGCggaattc	-4-
ggatccTGTCGTTTAAGTGGTTATGGCTAAGTGGCCCGAGGCGAATTC	6+
GGATCCTGTCGATAAGTGGAAGTACTTAAGGTAAAGTGAGGCggaattc	6+
ggatccTGTCGTTAAGTGAAGTCCGTTAAGTACCCGAGAGGCGAATTC	6+
GGATCCTGTCGTTAAGTGGACGAACATTAAGTGTTTTGAGGCGAATTC	7+
GGATCCTGTCGGCGAAGTGGTTAAGGTTTAAGTGGTTGAGGCGAATTC	7+
GGATCCTGTCGTTAAGTGGACGAACATTAAGTGTTTTGAGGCggaattc	7+
GGATCCTGTCGATAAGTGGGCTGAGTAAGTGGCGACAGAGGCGAATTC	5+
GAATTCGCCTCAAGTGGTTCGTCTAAGGACGTCTCTCGACAGGATCC	5+
ggatccTGTCGGTAAGTGCTTCGCGCTAAGTGACCTGAGGCGAATTC	6-
GAATTCGCCTCGTCACGATAAGTGGTGCTTGACACTTCGACAggatcc	6-
GGATCCTGTCGAGTACTTATTGTCGACTTAAGTGGTTGAGGCggaattc	7-
GGATCCTGTCGAGGTTAAGTGTTTTTTAATTGCTTGAGAGGCggaattc	7-
GGATCCTGTCGTGTTAAGTGGCCACTAGTGCCACTTAGAGGCGAATTC	9-
GGATCCTGTCGTGTTAAGTGGCCACTAGTGCCACTTAGAGGCggaattc	9-
GGATCCTGTCGTTAAGTGGTACTTGGAAGTTAAGTGGGAGGCGAATTC	10+
ATCCTGTCGGGAACTAAGTGGGCTAGAGTACTTAAGAGGCGAATTC	10+
GGATCCTGTCGTTAAGTGGTTTAGCGGCACTCCTTAAGAGGCGAATTC	10-
GAATTCGCCTCTTAAGGAGTGCCGCTAAACCACTTTAACGACAggatcc	10-
gaattccGCCTCCTCGAGTTAAGTGGAACTACTTACTCCGACAGGATCC	2-
GAATTCGCCTCACTTAAAGTGCCCTACGATAAGTGGGCGACAGGATCC	11-
GGATCCTGTCGTTAAGTGGTTTAGCGGTACTCCTTAAGAGGCGAATTC	14+

Figure 21. Tabular representation of base usage in XBap DNA binding sites.

A. The selected site for the XBap protein contains two motifs. Each of these motifs contains a core TAAG sequence. In addition there is one base upstream and three bases downstream of each core. This table indicates the distribution of bases upstream of each core and the downstream trios that they occur with.

B. Oligos designed for bandshift analysis on the basis of the above table. Nomenclature assigned principally on the basis of the first base.

T2 most commonly selected site

GR G Right

GL G Left

TR T Right

TL T Left

A

XTAAGXXX---XTAAGXXX					
MAIN SITE			SECOND SITE		
1st Base	Final trio	Total with Trio	1st Base	Final Trio	Total with Trio
60 T 11 C 6 A 2 G	TGG	79	6 G 6 C 5 T 1 A	TGG	18
7 T 1 G	TGC	8	10 T 2 G 1 A	TAC	13
3 T 2 A 1 C	TGT	6	6 T 1 C	TGC	7
1 T	TGA	1	6 T	AGG	6
1 T	CGG	1	3 T	TAG	3
1 T	GAG	1	2 T	TGT	2
1A	AGG	1	2 T	GAG	2
			1T	TGA	1
			1T	TGC	1
			1T	TAT	1
			1T	ATA	1
			1C	GAC	1

Total of Each Base
73 T
12 C
9 A
3 G

Total of Each Base
39 T
8 C
8 G
2 A

B T2 TTAAGTGG---TTAAGTGG

GL GTAAGTGG---TTAAGTGG

GR TTAAGTGG---GTAAGTGG

TR TTAAGTGG---TTAAGTAC

TL TTAAGTAC---TTAAGTGG

site is capable of forming a specific complex with XBap protein. Lane 1 of figure 22 indicates the complexes formed between the T2 probe and *in vitro* translated XBap protein. The specific complex is identified by the retardation of the complex by the addition of 9E10 antibody to the binding reaction (supershift, lane 5). The upper band clearly visible in lane 1 fails to supershift. In addition, unlabelled T2 oligonucleotide competes efficiently for T2 probe binding at both 10 and 100 fold molar excess of competitor (lane 2 and 3). Unlabelled oligonucleotide of a non-specific sequence competes less efficiently for binding when present at 100 fold molar excess (lane 4). As a control the oligonucleotide was also incubated with protein from an *in vitro* translation reaction to which no RNA was added (lanes 6-11). Thus, any complexes formed with this product are due entirely to proteins present in the reticulocyte lysate (note that the upper band present in lanes 1-5 is present at equal levels in lanes 6-11). As this is subject to variation between batches of reticulocyte lysate all subsequent bandshifts were performed with 'un-primed' controls using the same reticulocyte lysate used to prepare that batch of XBap.

Figure 23 demonstrates that both full length XBap (XB-F) and the homeodomain of XBap alone (XB-HD) can form a DNA/protein complex with the T2 site. A full experiment which demonstrates that the specificity of the homeodomain binding is the same as the full length is described in detail in chapter 5 but from this stage the experiments described may use either full length or homeodomain alone proteins.

In order to determine if XBap recognises the T2 oligonucleotide as effectively as it binds to other sequences selected the next two most commonly selected second sites were used as bandshift probes and as competitors to binding with T2 probe (Fig. 24). Lanes 4-11 demonstrate that if T2 oligonucleotide is used as probe none of the other sites can compete, at 10X or 100X molar excess of unlabelled oligonucleotide, more effectively the T2 oligonucleotide. Although all of the alternative sites form complex well with XBap (Lanes 11,17,22 and 27), T2 is equally as effective in competition for binding as

Figure 22. Bandshift to identify specific XBap Protein - DNA complex.

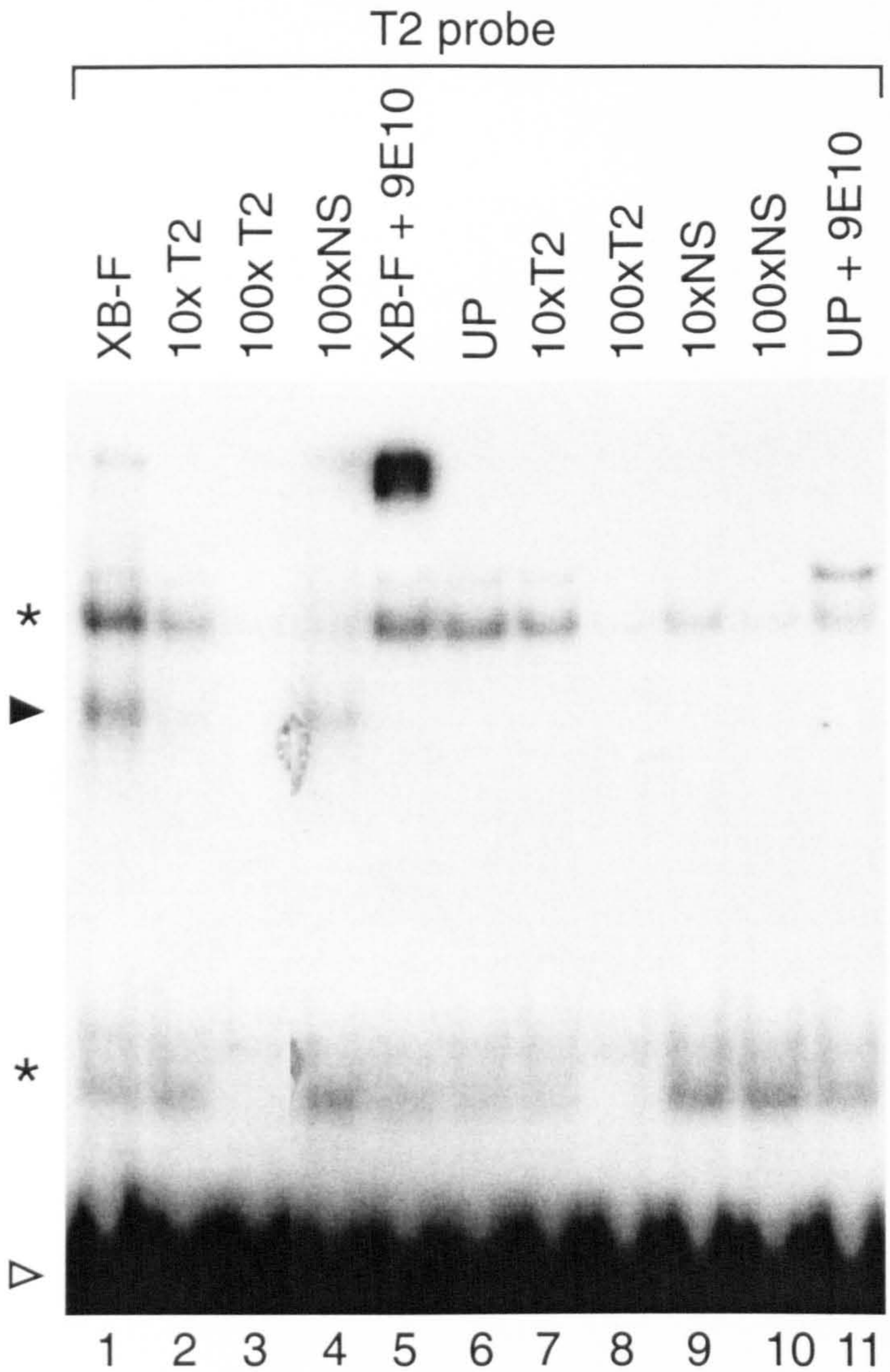
Lanes 1-11 all used wild-type oligo as probe. Lanes 2-4 show competition for binding with wild-type and non-specific oligos. In lane 5 9E10 antibody was used to supershift the XBap specific complex. Lanes 6-11 are a repeat of lanes 1-5 but using the H_20 control product. These lanes identify those complexes formed by proteins present in the reticulocyte lysate.

H_20 - *in vitro* translation control product.

Full - Full length XBap protein produced by *in vitro* translation of synthetic RNA.

T2 TTAAGTGG---TTAAGTGG

NS AGGGTTATTTTTAGAGCG



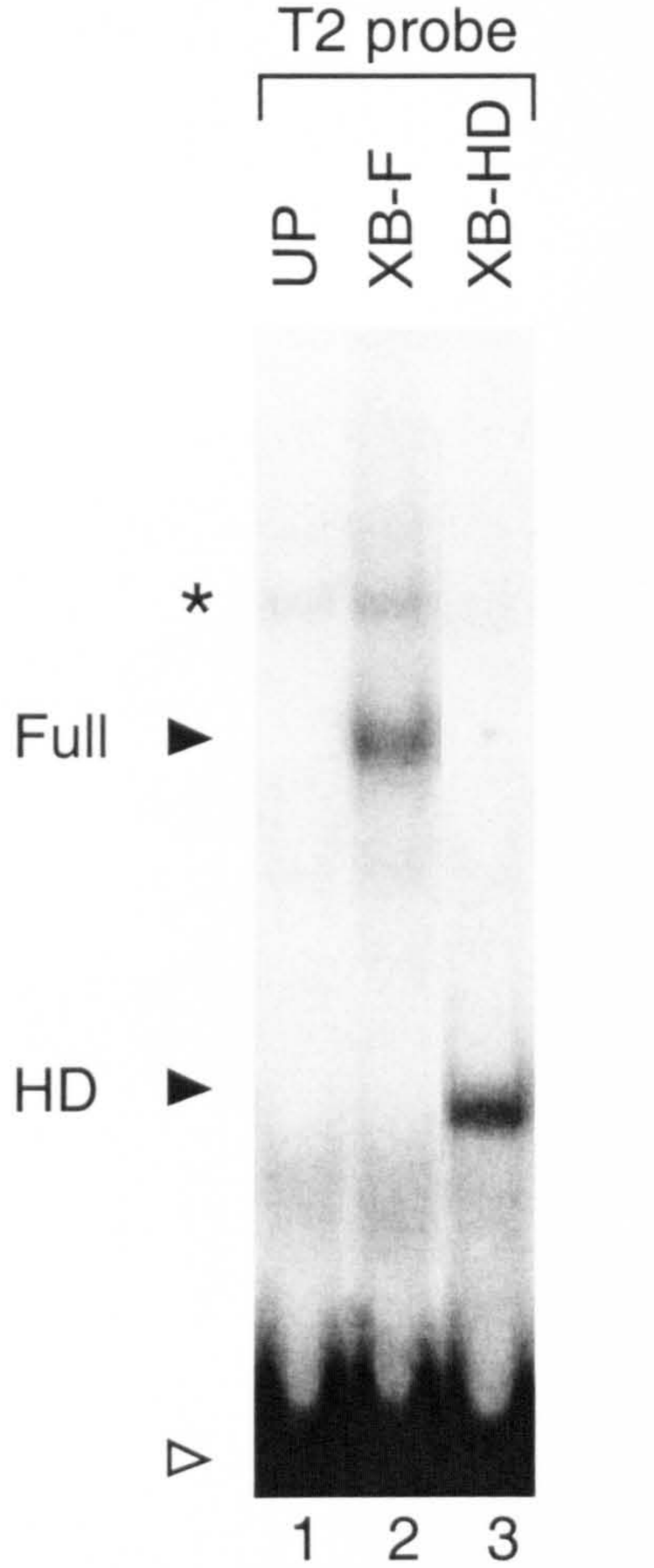
► = specific complex

▷ = free probe

* = background

Figure 23. Bandshift to demonstrate that both full length XBap protein and the XBap homeodomain alone can form complex with the T2 site

XB-F	Full length XBap protein produced by *in vitro* translation of RNA synthesised from construct ORF/Myc/pT7
XB-HD	The homeodomain of XBap produced by *in vitro* translation of RNA synthesised from construct HD/Ha/pT7



► = specific complex

▷ = free probe

* = background

Figure 24. Bandshift to compare the binding ability of the five most commonly selected sites.

The same XBap protein produced by *in vitro* transcription/translation was used in all of the lanes.

Lanes 1-11 used probe produced from the T2 oligo.

Lanes 12-16 used probe produced from the GR oligo.

Lanes 17-21 used probe produced from the GL oligo.

Lanes 22-26 used probe produced from the TR oligo.

Lanes 27-31 used probe produced from the TL oligo.

Each probe was tested for competition with T2 oligo at 10X and 100X levels and with 10X and 100X oligo corresponding to the specific probe.

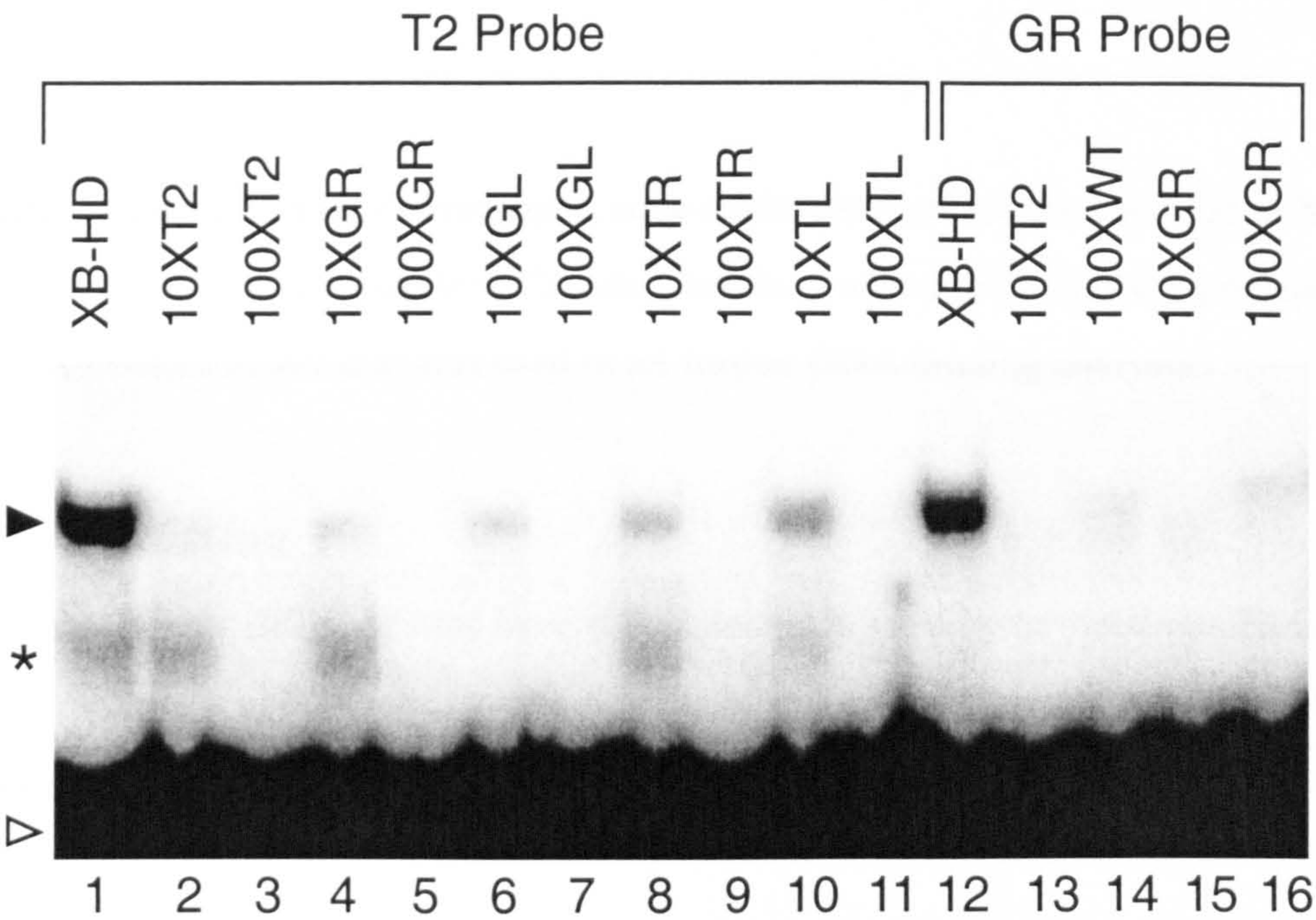
T2 TTAAGTGG---TTAAGTGG

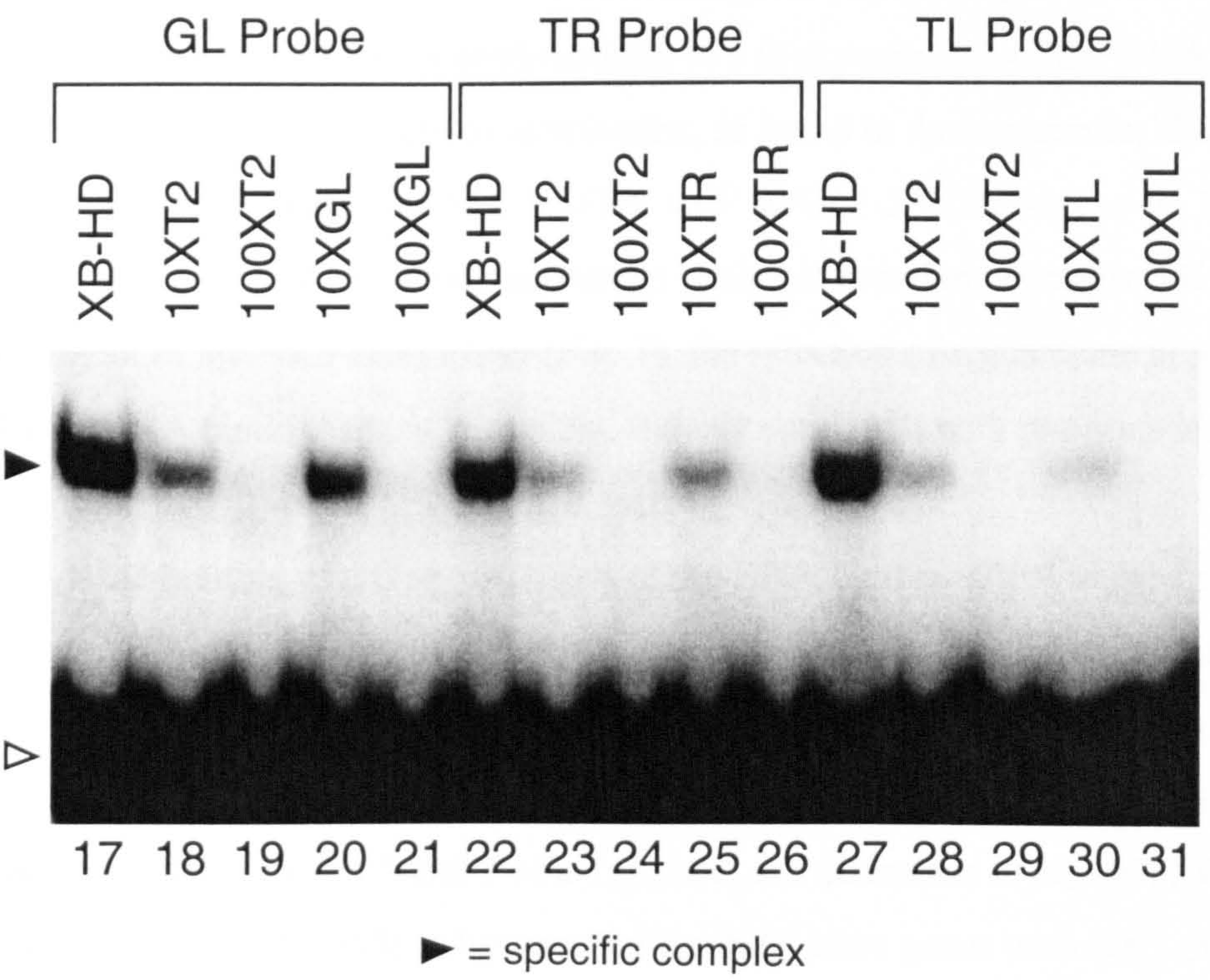
GL GTAAGTGG---TTAAGTGG

GR TTAAGTGG---GTAAGTGG

TR TTAAGTGG---TTAAGTAC

TL TTAAGTAC---TTAAGTGG





► = specific complex

▷ = free probe

* = background

the oligonucleotide itself. In summary, none of the alternative sites can bind to XBap protein more efficiently than the 'T2' site, therefore, as it demonstrated representative binding, it was this site that was used in all further DNA binding analysis.

4.3 Discussion

Given that so few different sites have been selected *in vitro* by homeodomains across divergent classes of homeobox genes it is especially surprising that within the NK class of homeobox genes that there should be two different *in vitro* DNA binding specificities.

Damante et 1996, and Pellizari *et al*, 1997 demonstrated that amino acid 54 is responsible for specification of base pair 4 of a DNA binding site ($C_1A_2A_3G_4$). TTF-1 (Nkx-2.1), with a tyrosine at position 54, binds preferentially to a G at position 4 of the DNA binding site, but mutation of this amino acid to methionine, as found in Antennapedia, changes the DNA binding preference from 5'-CAAG-3' to 5'-CAAT-3' (Damante *et al*, 1996). Tyrosine at position 54 of the homeodomain is almost invariant amongst NK class genes, including all of the NK3 class genes (Fig. 7), the selection of a guanosine at position 4 of the XBap DNA binding site is, therefore, entirely consistent with previous studies.

Although the binding of a G at position 4 of the DNA binding site was predictable, amino acids responsible for binding to position 1 have not yet been so clearly identified and it is, therefore, impossible to predict the identity of the base bound at this position. However, amino acids in the N-terminal arm have been implicated by several groups in the interaction with base pairs 1 and 2 of a $T_1A_2A_3T_4$ site (Damante *et al*, 1996, Wilson *et al*, 1996 and Hirsch *et al*, 1995). Alignment of the NK3 class genes with NK2 class genes demonstrates that the N-terminal arm is the region of most variability between the two classes. However, the binding of XBap to oligonucleotides containing a CAAG---CAAG motif must be studied in order to ascertain that the selection of a T at position 1 of the binding motif is a genuine requirement of XBap for optimal DNA binding.

While XBap recognises the T2 site *in vitro*, binding specificity may vary *in vivo* under the influence of other promoter elements and interactions with other proteins. This point is particularly relevant as XBap is a homeodomain protein and interaction with other proteins is a mechanism that has been shown to be used by certain classes of homeodomain proteins to achieve greater DNA binding specificity.

In the yeast Saccharomyces cerevisiae, cell type is determined by three key regulatory proteins, α1, α2, and a1, which are homeodomain-like proteins (helix-turn-helix proteins). These regulators act in different combinations to establish patterns of gene expression which result in distinct cell-type specific phenotypes. The yeast α2 protein represses transcription of the a-specific genes by binding to a 32bp sequence, the α2 site, located upstream of each target gene. Goutte *et al*, demonstrated that the a1 protein alters the binding specificity of the α2 protein such that it no longer binds with high affinity to the α2 site but instead acquires the ability to bind to a 29bp sequence, the a1- α2 site, located upstream of the haploid-specific genes. Although a1 alone binds DNA weakly and nonspecifically and α2 has only modest affinity and specificity for DNA the a1/ α2 heterodimer binds to its target site with high affinity and specificity (Goutte *et al*, 1993). This is the most striking example to date that interacting with other proteins can modulate the DNA binding characteristics of homeodomain type proteins.

One family of co-factors, the homeoproteins encoded by the *Drosophila extradenticle* and vertebrate *pbx* genes (together, the PBC genes), bind co-operatively to DNA with other homeoproteins and are important for Hox function *in vivo* (van Dijk and Murre, 1994, Chang *et al*, 1996 and Chan *et al*, 1997). One of the best characterised Hox-PBC binding sites is present in a 20bp oligonucleotide, repeat 3, which was identified in the 5' promoter region of the mouse *Hoxb-1* gene (Popperl *et al*, 1995). Hoxb-1 protein or its *Drosophila* ortholog *Labial* are both able to bind co-operatively with Exd to this binding site whereas other Hox proteins, such as Ubx or Hoxb-4 cannot. Furthermore, when lacZ reporter constructs containing three tandem copies of repeat 3 were introduced into either mouse or *Drosophila* embryos, *LacZ* expression patterns resembling the endogenous

Hoxb-1 or *labial* expression patterns were generated, respectively (Popperl *et al*, 1995, Chan *et al*, 1996, Chan *et al*, 1997). In *Drosophila* expression driven by this reporter gene, *3Xrpt3-LacZ*, requires both *exd* and *labial* functions. Thus, in both mouse and *Drosophila,* repeat 3 behaves as an exd-dependent Hoxb1-/labial autoregulatory enhancer element (Chan *et al*, 1997 and references therein).

A bipartite 10bp Hox-PBC consensus binding site 5'*TGATNN*AT[g/t][g/a], has been defined *in vitro* in which the PBC and Hox half sites, indicated by italics and underlining, respectively, overlap (Chang *et al*, 1996, Lu and Kamps, 1996). The central two basepairs (NN), which are predicted to contact the Hox N-terminal arm, have been shown to influence which Hox partner is incorporated into the heterodimer *in vitro* (Chan and Mann, 1996, Chang *et al*, 1996). In the repeat 3 binding site these two specificity-determining basepairs are GG. Changing these central basepairs to TA results in an oligonucleotide that generates a *deformed* expression pattern instead of a *labial* expression pattern *in vivo* (Chan *et al*, 1997). Similar binding sites to these are present in the *Drosophila deformed* gene and in the mouse ortholog, Hoxb-4. These binding sites are able to generate *deformed* or *Hoxb-4* expression patterns in *Drosophila* or mouse embryos, respectively. These results demonstrate that PBC proteins co-operate with multiple Hox proteins *in vivo*. Further, in agreement with *in vitro* studies, they illustrate the importance of the two central basepairs in the Hox-PBX binding sites in distinguishing between different Hox specificities *in vivo* (Chan *et al*, 1997).

The Hox proteins contain a conserved YPWM motif N-Terminal of the homeodomain that, together with the N-terminal arm mediates Hox/PBC interaction (Shanmugam *et al*, 1997). This motif is not present in the NK Class genes but the hydrophobic nature of the NK2 box downstream of the homeobox has lead to the suggestion that it may mediate protein-protein interactions (Chen *et al*, 1995).

If it is the case that interacting proteins modify the DNA binding specificity of homeoproteins *in vivo* is there anything to be gained from *in vitro* binding site selection

studies? The examples described in the chapter introduction, TTF-1, Tinman and CBF, demonstrate that *in vitro* selected DNA binding sites can be representative of binding of those proteins to regulatory elements *in vivo*. In addition, although interacting proteins may modify the binding sites of homeoproteins they are not necessarily changed beyond recognition. The consensus Pbx/Hox heterodimer binding site, 5'-TGATNNAT[g/t][g/a] (an example of which has been identified *in vivo* in the 5' promoter region of Hoxb-1) is produced by an overlap of the Pbx and Hox binding sites. The monomer sites would, therefore, still be useful when screening potential targets for binding sites (Chang *et al*, 1996).

In summary, data to date supports *in vitro* binding site selection studies as a good starting point for identifying the basis of a potential *in vivo* DNA binding site, however, further studies are necessary to identify the additional factors which contribute to the subtleties of each homeoproteins *in vivo* specificity.

Chapter 5

Characterisation of the DNA binding of XBap

5.0 Chapter 5: Characterisation of the DNA binding of XBap

5.1 Introduction

The previous chapter demonstrated that XBap protein can form a specific complex with the most representative *in vitro* selected site but the region of the protein responsible for this specificity and which features of the site are required for optimal DNA binding needed to be determined. A comparison of the binding characteristics of XBap with other members of the class may provide further clues as to the mechanism of specificity.

Although the EMSA analysis indicated that XBap protein can form a specific complex with the T2 site it did not address whether the two motifs and the 5'-TAAG-3' core are actually required for optimal DNA binding. The simplest way to address this question is to compare the ability of XBap protein to bind to variant oligonucleotide sequences *in vitro*.

In the majority of cases examined, including TTF-1, the DNA binding specificity of HD-containing protein resides exclusively in the HD itself (Damante *et al*, 1994). In some cases, however, other domains including POU domains, leucine zippers, LIM domains and zinc-fingers, are able to co-operate with the HD in determining the DNA binding specificity of the entire protein. Whether this is true for XBap can be tested using *in vitro* translated protein produced from a truncated XBap construct, which encodes only the homeodomain, in an EMSA assay.

XBap is just one member of a growing family of *NK3* class genes. Two NK class genes have been identified in mouse and humans. *Bapx1* is identical across the homeodomain with *bagpipe* and in mouse is the earliest developmental marker of the presclerotome portion of the somite and of the gut mesentery. At later stages *Bapx1* is expressed in all cartilaginous condensations which will subsequently undergo endochondral bone formation (Triboli *et al*, 1997). *Nkx-3.1* expression in the male urogenital sinus, prostatic buds and testis suggests that it may play a regulatory role in the development of the male urogenital system of mouse and human (Sciavolino *et al*, 1997 and He *et al*, 1997). In

order to determine if the XBap DNA binding characteristics are representative of the class it would be informative to carry out the same EMSA analysis with Drosophila Bagpipe and the mouse proteins Bapx1 and Nkx-3.1. If any differences arise site-directed-mutagenesis may be used to pinpoint the amino acids responsible.

5.2 Results

5.2.1 EMSA assays

Synthesis of probes for the EMSA analyses were all performed under the same conditions according to the same protocol. The binding of each mutant probe was examined at least twice with at least two independent probe preparations. Oligonucleotides were annealed followed by a 3' fill-in reaction with Klenow in which radiolabelled CTP was incorporated. Probe was purified using D50 Spin Columns (Kodak, UK). The amount of probe used in each reaction was determined by the cpm (20,000 cpm / reaction). Binding reactions were carried out at RT (below 24° C) for 15 minutes in the absence of probe with a further 15 minute RT incubation on addition of probe.

Although there was no exact quantification of probe the yield was assumed to be 50% and consistency was monitored by the total cpm of each synthesis. In every experiment the effect of each oligonucleotide was observed not only by the amount of probe/protein complex formed but also by the comparative ability of the oligonucleotide as an unlabelled oligonucleotide competitor.

5.2.2 Single versus double motif analysis

Why does optimal binding of XBap require two copies of the motif? There are three possible explanations; either XBap forms dimers in solution and binds to DNA as a dimer, or the binding of one XBap protein to DNA increases the binding of a second protein (co-operative binding), or just one XBap protein binds to the DNA binding site but the two motifs are essential in order to achieve optimal stability.

To study whether two motifs are required for optimal DNA binding by XBap two mutant oligonucleotides were designed. In the first, the core of the right-hand motif has been mutated (single mutant or SMut), and in the second, both cores have been mutated (double-mutant or DMut). The mutation changes the core of the motif from 5'-TAAG-3' to 5'-CGAT-3'.

Figure 25 demonstrates that the unlabelled double mutant is unable, at either 10 or 100 fold molar excess, to compete with T2 radio-labelled probe for binding to XBap (compare lanes 7 and 8 with 3 and 4). No complex was detectable when using the double mutant as probe (data not shown). This indicates that the double-mutant oligonucleotide is unable to bind XBap and that XBap specifically recognises the T2 sequence.

The single-mutant oligonucleotide contains just one functional copy of the binding motif and is shown in lanes 5 and 6 to be less effective at competing for XBap binding than the T2 oligonucleotide (lanes 3 and 4). In addition, if the single-mutant oligonucleotide is used as probe the complex is barely visible (lane 10). These results indicate that two motifs allow optimal DNA binding of XBap.

The determination that two copies of the binding motif are required for optimal DNA binding of XBap raises more issues. The two motifs of a direct repeat may not be equivalent in their.contribution to binding and it is also possible that XBap may bind to a palindromic arrangement of motifs. Two new oligonucleotides were synthesised in order to address these questions; SMutL, which has the left hand core motif mutated instead of the right, and PAL in which the two motifs are in palindromic format.

efficiently

Lanes 3-6 of figure 26 demonstrate that both SMutR and SMutL can only compete for binding at 100 fold molar excess with T2 oligonucleotide, although SMutR appears marginally more effective. Compared with T2, XBap binds slightly less efficiently with both SMutR (lane 9) or SMutL (lane 10) revealing that both sites appear to be of equivalent importance in the optimal binding of XBap. Lanes 7 and 8 of figure 26 show that the

Figure 25. Bandshift to study the requirement of XBap for two copies of the selected core for DNA binding.

Lanes 1-8 all used wild-type oligo as probe with wild-type, single-mutant or double-mutant oligos as probe. Lanes 9 and 10 used single mutant oligo as probe.

H_20 - *in vitro* translation control product.

Full - Full length XBap protein produced by *in vitro* translation of synthetic RNA.

T2 TTAAGTGG---TTAAGTGG

SMut TTAAGTGG---TTCGATGG

DMut TTCGATGG---TTCGATGG

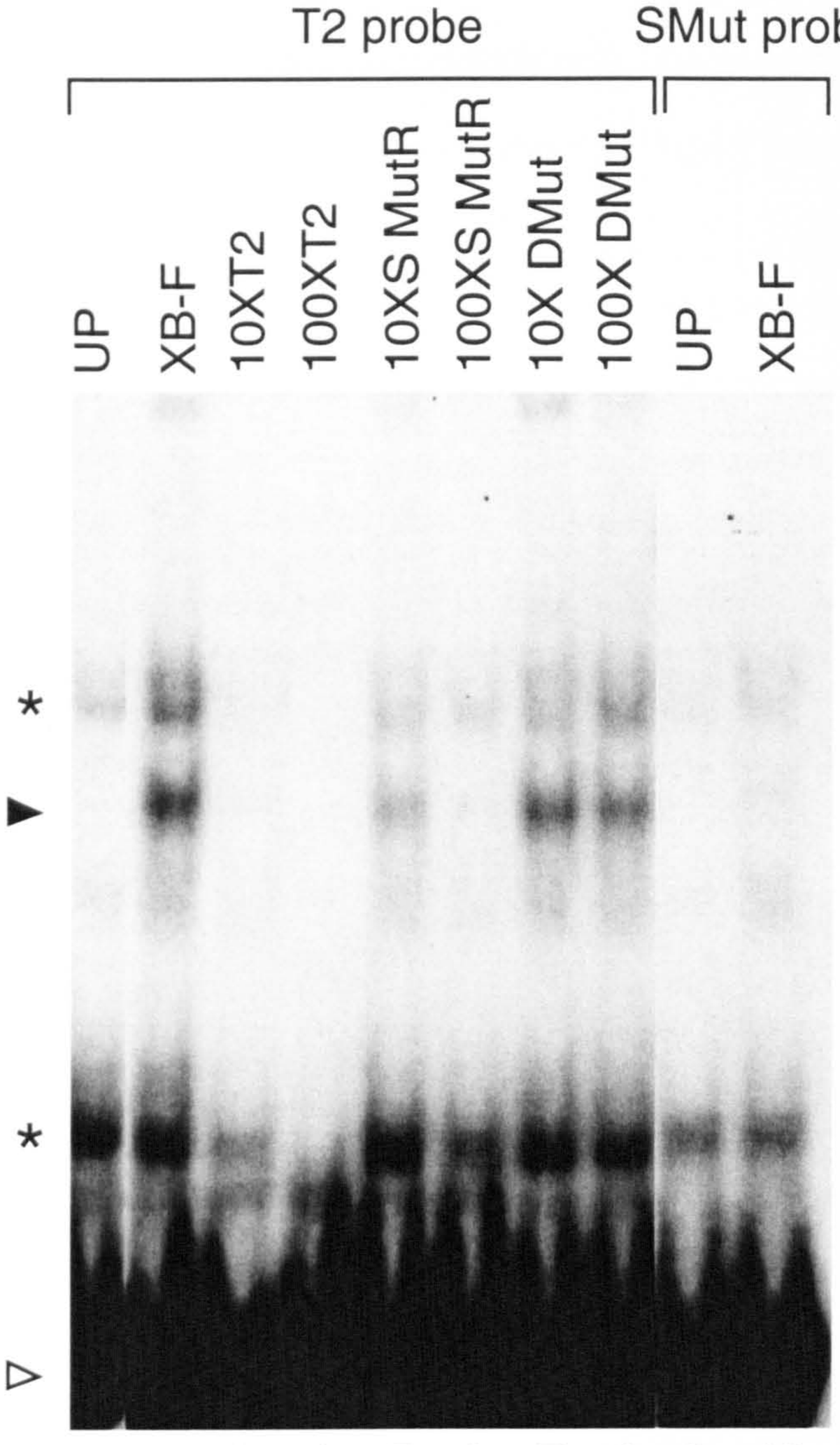

T2 probe
SMut probe
UP
XB-F
10XT2
100XT2
10XS MutR
100XS MutR
10X DMut
100X DMut
UP
XB-F
*
*
1 2 3 4 5 6 7 8 9 10
► = specific complex
▷ = free probe
* = background

Figure 26. Bandshift to study the ability of XBap to bind to asymmetrically mutated binding sites and palindromic binding sites.

Lanes 1-8 all used T2 oligo as probe with single-mutant right, single-mutant left, and palindrome oligos as competitors.

Lane 9 used single-mutant right as probe, lane 10 as single-mutant left and lane 11 the palindrome oligo.

H_20 - *in vitro* translation control product.

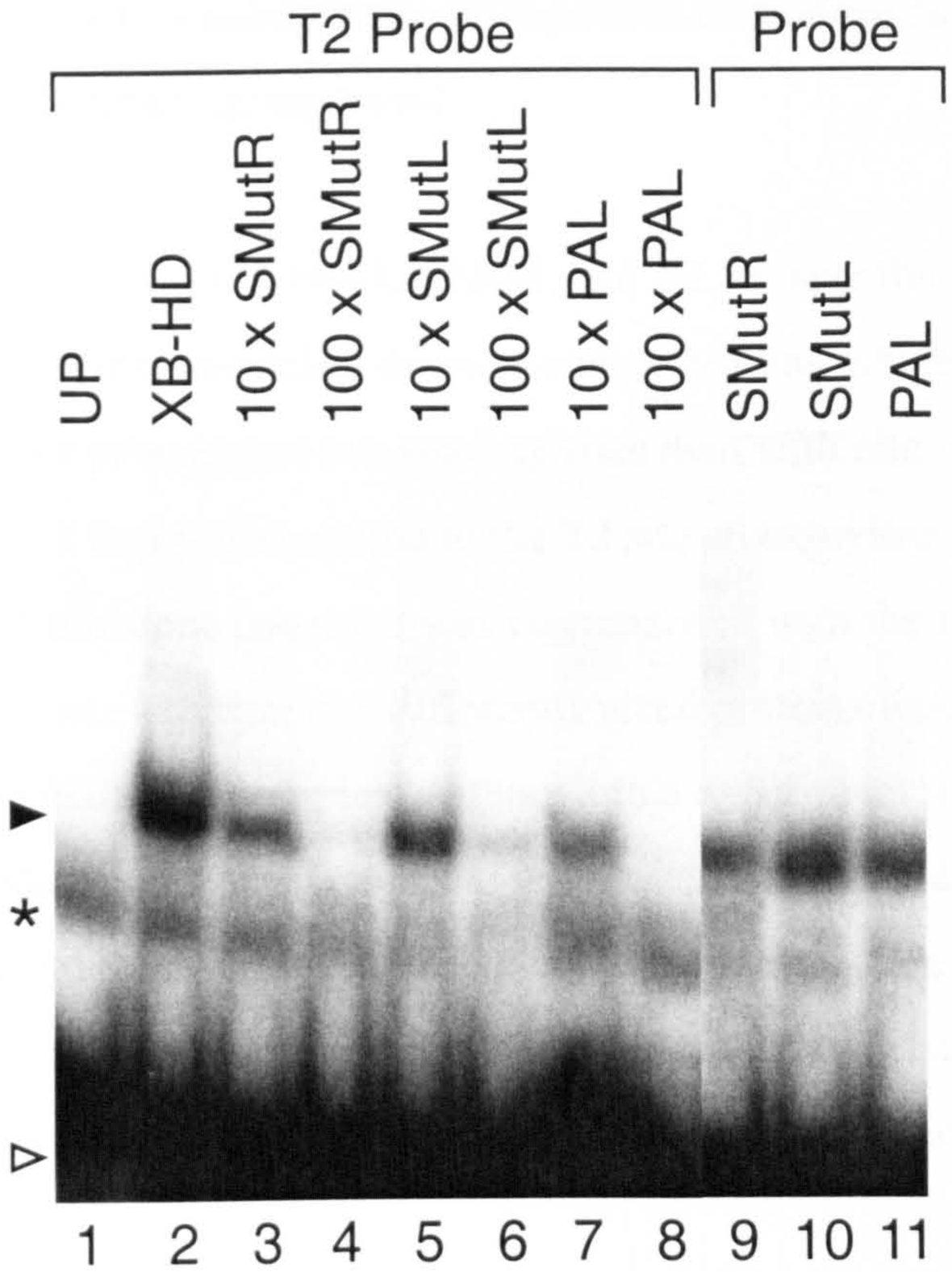
Full - Full length XBap protein produced by *in vitro* translation of synthetic RNA.

T2 TTAAGTGG---TTAAGTGG

SMutR TTAAGTGG---TTCGATGG

SMutL TTCGATGG---TTAAGTGG

PAL TTAAGTGG---CCACTTAA



► = specific complex
▷ = probe
* = background

palindromic arrangement of motifs can only compete as efficiently as the single mutant probes. Lane 11 further reveals that complex can form with PAL, however, although XBap can bind weakly to a palindromic arrangement of binding motifs, DNA binding is optimal with a direct repeat arrangement.

Complexes formed with PAL, SMutR, SMutL and T2 all have the same mobility, therefore, eliminating the possibility of co-operative binding which would result in a larger complex in the presence of two binding sites than with one. In order to address whether one or two XBap proteins bind to the T2 site an experiment was carried out in which a homeodomain alone construct was co-translated with the full-length XBap construct; heterodimers between two differently sized proteins would be expected to produce an intermediate size complex. Although this experiment was attempted with DNA or RNA as starting material and even with *in-vitro* translated protein being mixed in the binding reaction, no evidence was ever seen of an intermediate size complex (data not shown). There was also no intermediate complex produced by a bandshift reaction with XBap homeodomain and *Drosophila Bagpipe* protein (data not shown).

5.2.3 5'-TAAG-3' Core versus 5'-CAAG-3' Core Analysis

To study whether XBap specifically needs a 5'-TAAG-3' core for optimal DNA binding two mutant oligonucleotides were designed in which either the core of one (C1), or both motifs (C2), were mutated to 5'-CAAG-3' (Fig. 27).

Although the C1 oligonucleotide can compete for XBap binding with T2 probe (lanes 5 and 6) it is not as efficient as T2 (lanes 3 and 4) and the C2 oligonucleotide is an even less efficient competitor. If the C1 oligonucleotide is used as probe complex formation is very poor and if C2 oligonucleotide is used as probe no complex is detected at all (lane 12). These results indicate that a sequence with *two* 5'-TAAG-3' cores is preferred for optimal XBap binding and that 5'-TAAG-3' is an absolute requirement

Figure 27. Bandshift to study the requirement of XBap for a TAAG core, rather than a CAAG core, for DNA binding.

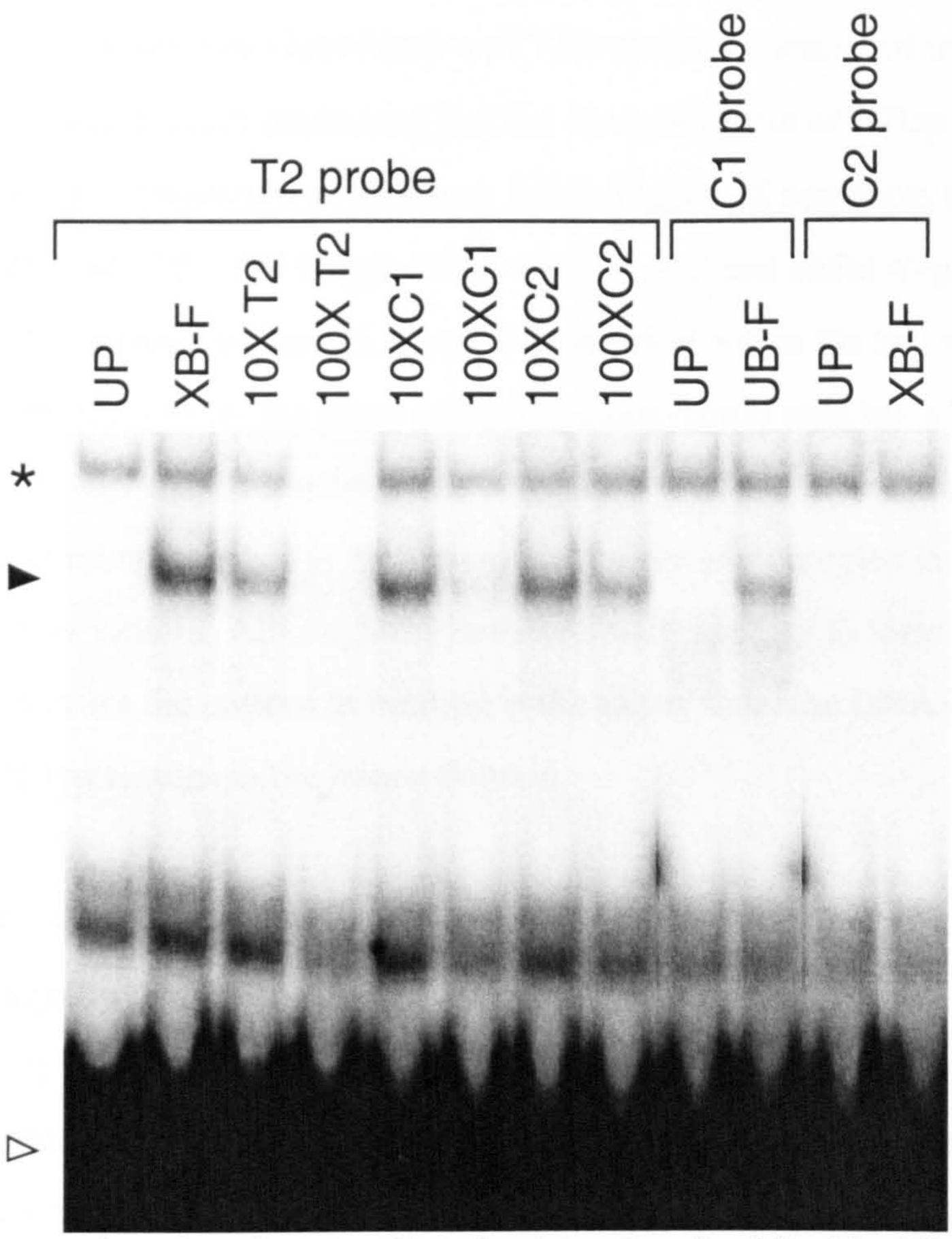
Lanes 1-8 all used T2 oligo as probe. Lanes 3-8 used T2, C1 mutant or C2 mutant as competitors for binding to full length XBap. Lanes 9 and 10 used the C1 mutant as probe and 11 and 12 used C2 oligo as probe.

H_20 - *in vitro* translation control product.

Full - Full length XBap protein produced by *in vitro* translation of synthetic RNA.

T2	TTAAGTGG---TTAAGTGG
C1	TTAAGTGG---TCAAGTGG
C2	TCAAGTGG---TCAGGTGG

.


T2 probe
C1 probe
C2 probe
UP
XB-F
10X T2
100X T2
10XC1
100XC1
10XC2
100XC2
UP
UB-F
UP
XB-F
*
▶
▷
1 2 3 4 5 6 7 8 9 10 11 12
▶ = specific complex
▷ = free probe
* = background

5.2.4 DNA binding characteristics of the XBap homeodomain.

The homeodomain alone was cloned into a pT7TS vector as described in the materials and methods section. Protein containing just the homeodomain of XBap (HD) was generated by *in vitro* translation of synthetic RNA. Figure 28 compares the behaviour of this protein with that of the full length XBap with C1, C2 and SMut oligonucleotides. Lanes 2,3 and 4 compared to lanes 5, 6 and 7 demonstrates that the full length and homeodomain proteins show the same pattern of competition with the SMut probe. Lanes 9 and 10 and lanes 15 and 16 illustrate that complex formation between both proteins and the C1 and SMut mutant probes is inefficient and barely any complex is formed with the C2 probe by either protein. Although the homeodomain appears to form complex more efficiently in all cases the pattern of binding is the same, thus, the DNA binding specificity of XBap resides in the homeodomain.

5.2.5 DNA binding characteristics of other NK3 class proteins

Drosophila Bagpipe

Drosophila Bagpipe was the first NK3 class protein to be identified and it was, therefore, of interest to determine whether the DNA binding characteristics had been conserved throughout evolution from *Drosophila* to *Xenopus*. In order to compare the binding of *Drosophila* Bagpipe with that of XBap the same series of mutant oligonucleotides were used in an EMSA study. Figure 29 illustrates that *Drosophila* Bagpipe can form complex efficiently with the T2 site (lane1), the absence of competition by non-specific oligonucleotide (lane 12 and 13) confirms that complex formation is sequence specific. Lanes 4-5 show that the single mutant oligonucleotide is unable to compete efficiently with the T2 probe for binding. Lane 14 shows that *Drosophila* Bagpipe complex formed with the single-mutant probe is barely detectable. Thus, *Drosophila* Bagpipe also needs two copies of the motif for optimal DNA binding. Lanes 8 and 9, however, show that competition with the C1 oligonucleotide is almost as effective as competition with the T2 oligonucleotide (lanes 2 and 3) and lanes 10 and 11 show that C2 is also able to compete with T2 although not as efficiently as C1. Furthermore, lanes 15 and 16 demonstrate that

Figure 28. Bandshift to compare the DNA binding properties of full length XBap protein and XBap homeodomain only protein.

Lanes 1-9 used the mutant oligos as probes. Each probe was used with H_2O control product, Full length XBap protein and XBap homeodomain. Lanes 11-16 compare the behavior of Full XBap and HD XBap with T2 oligo as probe and single-mutant oligo as competitor.

H_20 - *in vitro* translation control product.

Full - Full length XBap protein produced by *in vitro* translation of synthetic RNA.

HD - The homeodomain of XBap produced by *in vitro* translation of synthetic RNA.

T2	TTAAGTGG---TTAAGTGG
C1	TTAAGTGG---TCAAGTGG
C2	TCAAGTGG---TCAGGTGG
SMut	TTAAGTGG---TTCGATGG

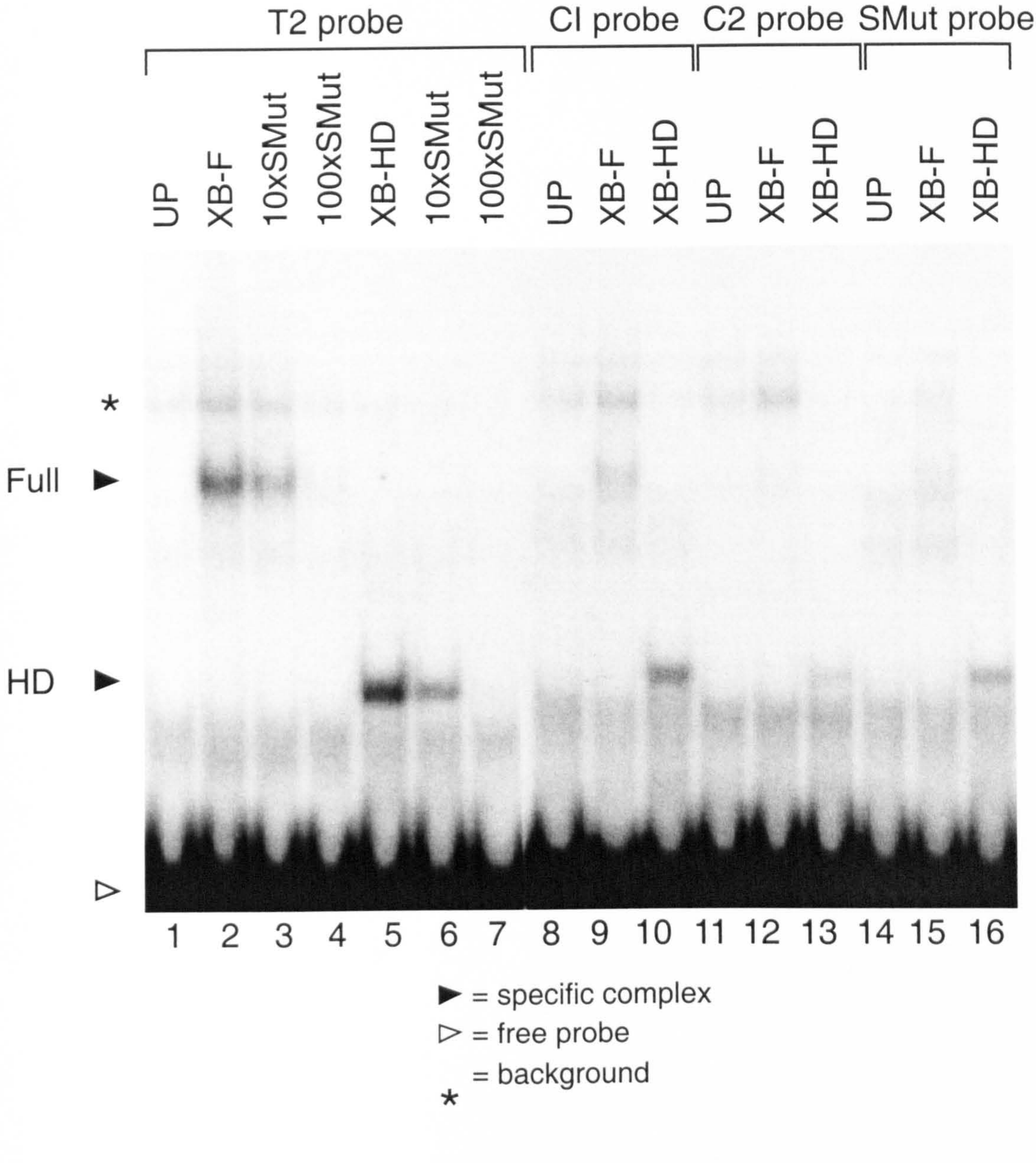

T2 probe
CI probe
C2 probe
SMut probe
UP
XB-F
10xSMut
100xSMut
XB-HD
10xSMut
100xSMut
UP
XB-F
XB-HD
UP
XB-F
XB-HD
UP
XB-F
XB-HD
*
Full
HD
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
= specific complex
= free probe
= background
*

Figure 29. Bandshift to study the DNA binding properties of *Drosophila* Bagpipe.

A

Panel A used T2 oligo as probe with T2, single-mutant, double-mutant, C1, C2 and non-specific oligos as competitors for binding to *Drosophila* Bagpipe protein.

B

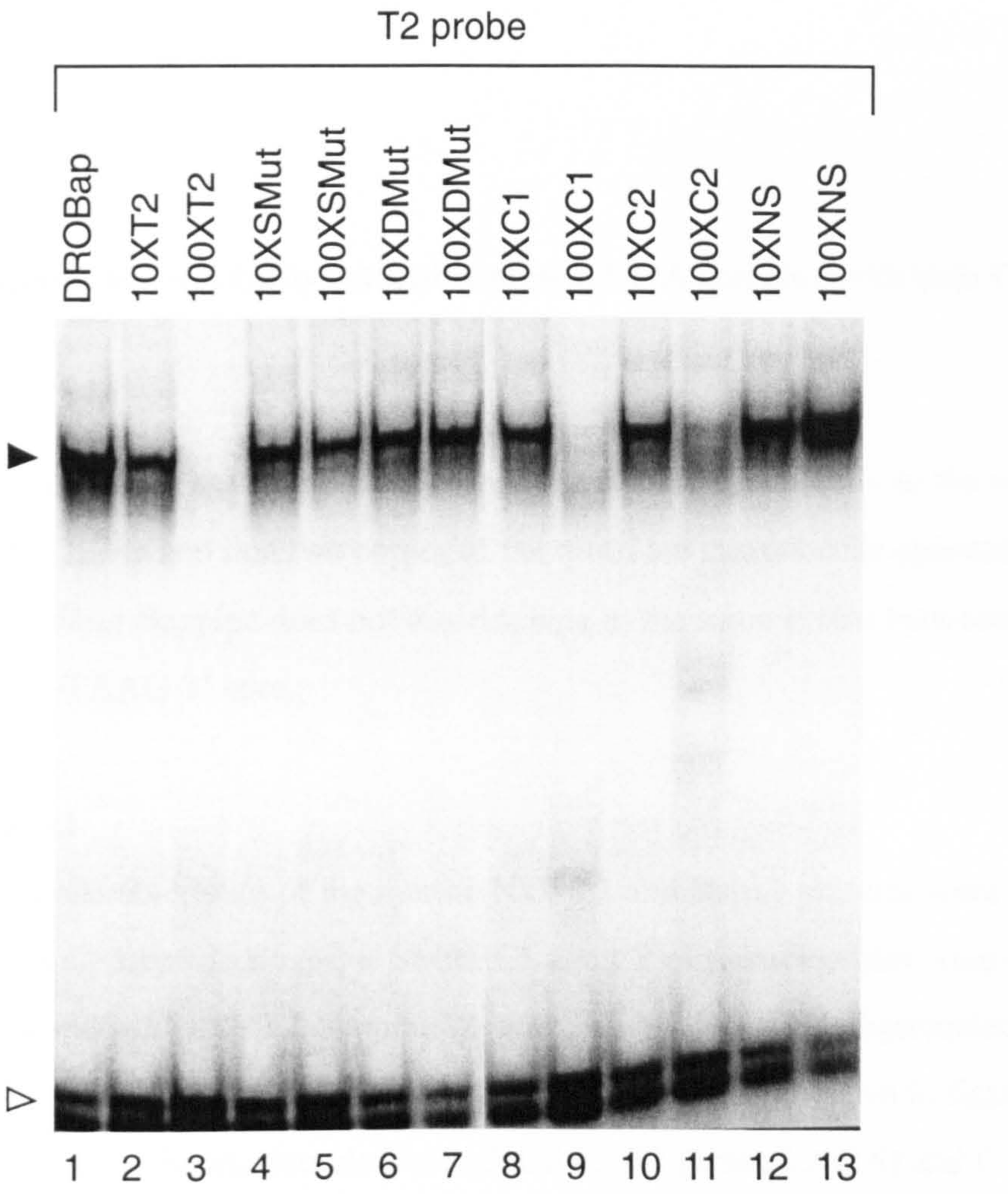
Panel B shows binding of *Drosophila* Bagpipe to single-mutant, C1 and C2 probes.

H_20 - *in vitro* translation control product.

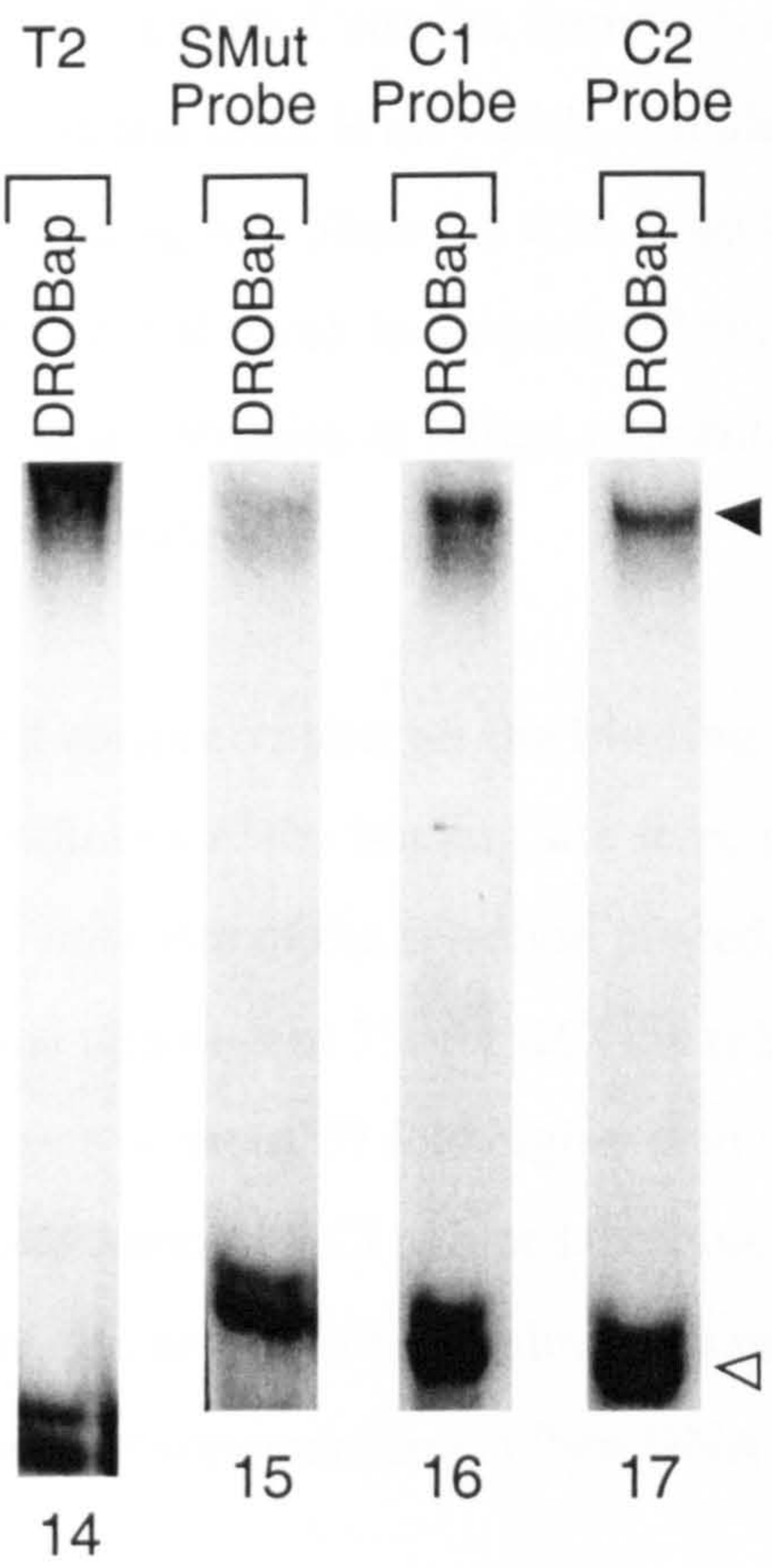
DROBap - *Drosophila* Bagpipe produced by purification of bacterially expressed protein

T2	TTAAGTGG---TTAAGTGG
SMut	TTAAGTGG---TTCGATGG
DMut	TTCGATGG---TTCGATGG
C1	TTAAGTGG---TCAAGTGG
C2	TCAAGTGG---TCAGGTGG
NS	AGGGTTATTTTTAGAGCG

A



B



► = specific complex

▷ = free probe

Drosophila Bagpipe can form easily detectable protein/DNA complex with both C1 and C2 oligonucleotides.

These experiments indicate that *Drosophila* Bagpipe can bind specifically to the same site as that selected for XBap and that two copies of the motif are essential for optimal binding but *Drosophila* Bagpipe does not discriminate to the same extent between a 5'-CAAG-3' and a 5'-TAAG-3' core.

Nkx-3.1 and Bapx1

The DNA binding characteristics of the murine Nkx-3.1 and Bapx1 proteins were studied in the same way as *Drosophila* Bagpipe. SMut, C1 and C2 oligonucleotides were used as competitors in comparison with T2 competition and SMut, C1 and C2 oligonucleotides were also used as probes. The results of this assay with Nkx-3.1 are shown in figure 30. Competition with SMut oligonucleotide (lanes 5 and 6), C1 (lanes 8 and 9) and C2 oligonucleotides (lanes 10 and 11) are not as efficient as competition with T2 oligonucleotide (lanes 3 and 4). Complex formation is poor with SMut probe (lane 13) and C1 probe (lane 15) and there is no visible complex formed with the C2 probe. An identical pattern of binding was observed with Bapx1 protein and each of the mutant oligonucleotides (data not shown). In summary, both Bapx1 and Nkx-3.1 illustrate the same DNA binding characteristics as XBap, requiring two motifs with a 5'-TAAG-3' core for optimal DNA binding.

5.2.6 Effect of salt concentration on the binding of XBap and *Dros* Bagpipe

The fact that the conditions of the binding site selection are not close to intracellular conditions may be a criticism of the selection procedure. In addition, studies by Damante *et al*, have shown that although at 75mM KCl the relative binding of TTF-1 with a specific DNA sequence is about 50 fold higher than with a non specific DNA sequence, increasing the concentration of KCl reduced the binding affinity of TTF-1 for both sequences. However, the non-specific binding was more sensitive than the specific binding to increasing salt concentration. When DNA binding reactions were performed at

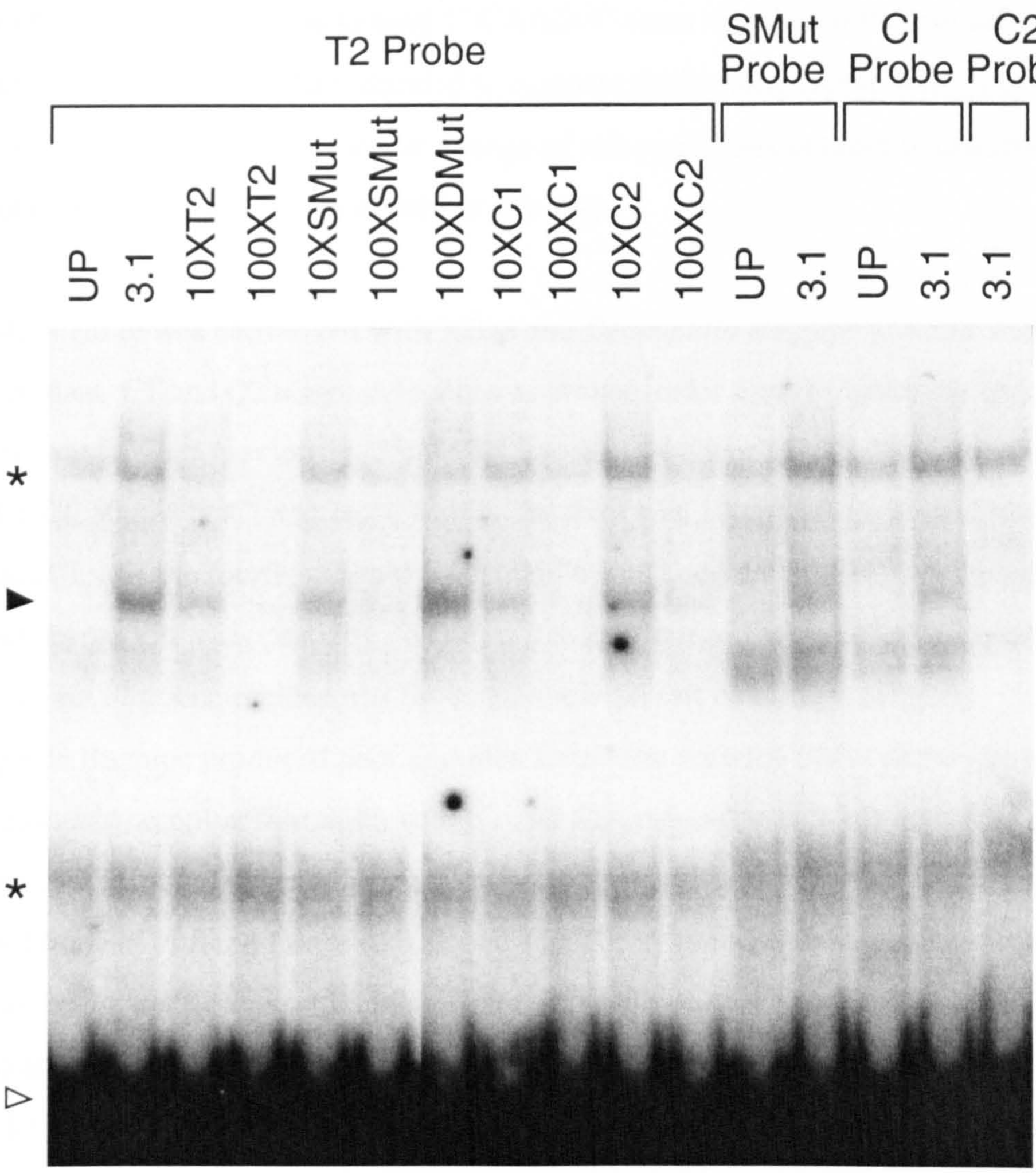
Figure 30. Bandshift to study the DNA binding properties of Nkx-3.1.

Lanes 1-11 used T2 oligo as probe with T2, single-mutant, double-mutant, C1 and C2 oligos as competitors for binding to Nkx-3.1 protein (produced by *in vitro* translation of synthetic RNA). Lanes 11-16 use single-mutant, C1 and C2 oligos as probes with Nkx-3.1 protein and H_2O control product.

H_20 - *in vitro* translation control product.

Full - Full length Nkx-3.1 protein produced by *in vitro* translation of synthetic RNA.

T2	TTAAGTGG---TTAAGTGG
SMut	TTAAGTGG---TTCGATGG
DMut	TTCGATGG---TTCGATGG
C1	TTAAGTGG---TCAAGTGG
C2	TCAAGTGG---TCAGGTGG



► = specific complex

▷ = free probe

* = background

temperature and salt conditions close to the mammalian intracellular environment TTF-1HD binds the specific sequence with an affinity at least 1000 fold higher with respect to the non-specific sequence (Damante *et al*, 1993). It was, therefore, possible that the ability of *Drosophila* Bagpipe to bind 5'-CAAG-3' cores was due to the low salt concentration. It was, therefore, decided to compare the binding characteristics of both XBap and *Drosophila* Bagpipe under a range of salt conditions in order to determine if the specificity of the proteins is altered in any way.

An EMSA assay was carried out with XBap and *Drosophila* Bagpipe proteins with T2, single-mutant, C1 and C2 oligonucleotides as probes under four different salt conditions. The first was as used previously, 75mM KCl with no NaCl or $MgCl_2$, the second was 75mM KCl, 50mM NaCl and 1mM $MgCl_2$, the third was 100mM KCl, 85mM NaCl and 2mM $MgCl_2$, and the fourth, 400mM KCl, 50mM NaCl and 2mM $MgCl_2$. Although complex formation was optimal with Buffer 2 the pattern of *Drosophila* Bagpipe binding to the variant oligonucleotides was the same under all salt conditions (Fig.31). *Drosophila* Bagpipe produced poor complex formation with the SMut probe (lanes 5-8) and reasonable complex formation with C1 and C2 probes (lanes 9-16) under the salt conditions of buffers 1-3 but binding to any oligonucleotide was barely detectable under the conditions of buffer 4 (lanes 4, 8,12 and 16). XBap complexes with T2 were formed under all buffer conditions, although buffer 1 (lane 17, as used in all previous binding assays) and buffer 2 (lane 18) allowed more complex formation than buffers 3 and 4 (lanes 19 and 20). XBap illustrated the same pattern of complex formation with all the mutant probes as observed previously under all of the salt conditions but the data is not shown as complexes formed with oligonucleotides other than T2 are barely detectable under all salt conditions except buffer 1.

5.2.7 Site directed mutagenesis

The comparative studies of the *Xenopus, Drosophila* and mouse NK3 family proteins reveal that all of the proteins require two copies of a motif but *Drosophila* Bagpipe

Figure 31. Bandshift to study the DNA binding of *Drosophila* Bagpipe to four different probes under a range of salt conditions.

Lanes 1-16 used *Drosophila* Bagpipe with the four probes studied previously, each with binding buffers B1-B4. Lanes 17-20 used XBap with T2 probe under the four buffer conditions.

B1 75mM KCl

B2 100mM KCl, 50mM NaCl, 1mM $MgCl_2$

B3 100mM KCl, 85mM NaCl, 2mM $MgCl_2$

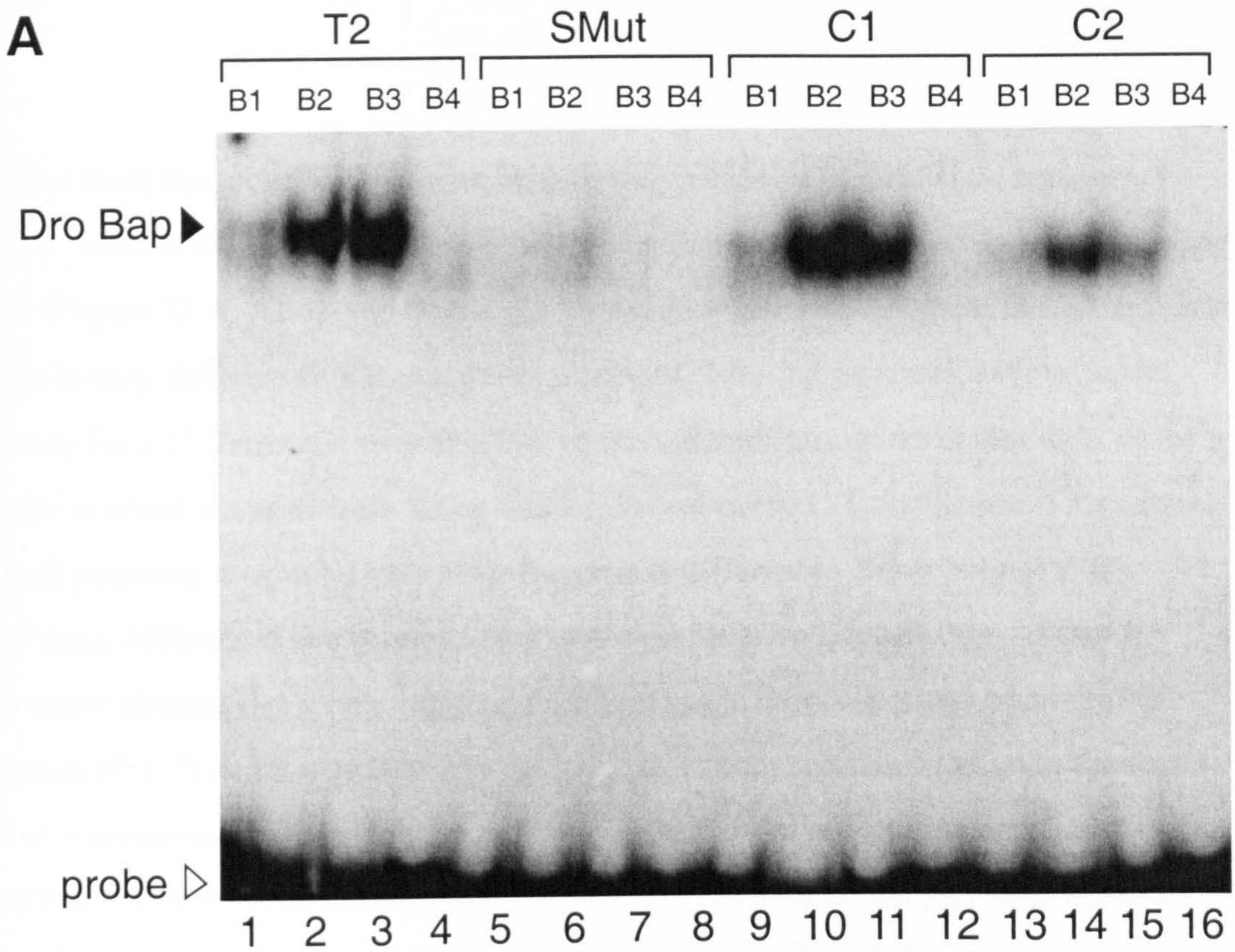
B4 400mM KCl, 50mM NaCl, 2mM $MgCl_2$.

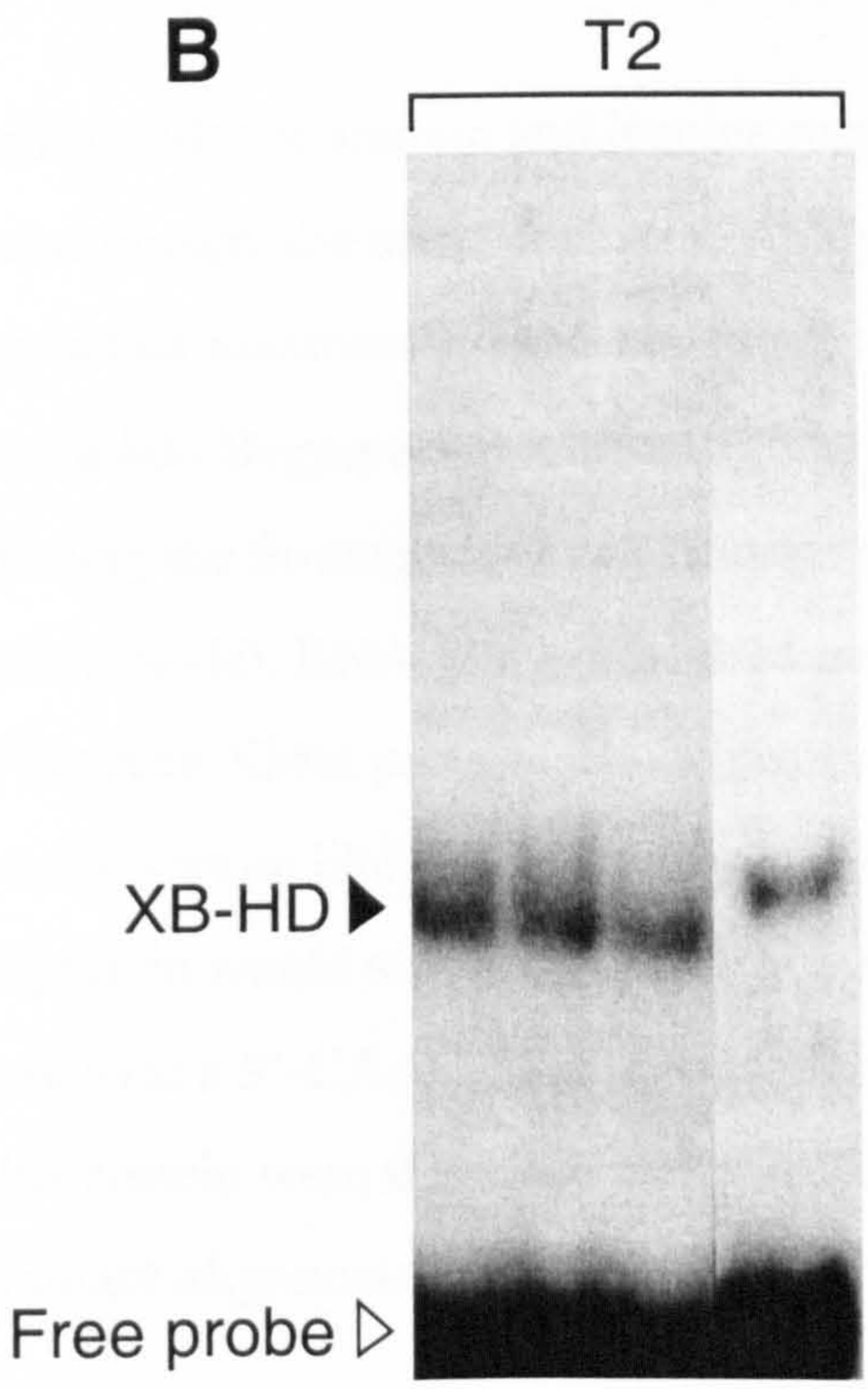
T2 TTAAGTGG---TTAAGTGG

SMut TTAAGTGG---TTCGATGG

C1 TTAAGTGG---TCAAGTGG

C2 TCAAGTGG---TCAGGTGG


A
T2
SMut
C1
C2
B1 B2 B3 B4 B1 B2 B3 B4 B1 B2 B3 B4 B1 B2 B3 B4
Dro Bap
probe
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16


B
T2
XB-HD
Free probe
17 18 19 20

stands out from the group in that it can bind equally well to a 5'-CAAG-3' core as a 5'-TAAG-3' core. If the amino acid sequences of the homeodomains of all these proteins are aligned (Figure 32.A. XBap and Bapx1 are identical) it becomes apparent that *Drosophila* Bagpipe is only different to XBap at seven positions. Nkx-3.1 demonstrates the same specificity for a 5'-TAAG-3' core as XBap so the assumption can made that none of the positions at which it varies from XBap will be critical to the C/T distinction. This leaves just three positions at which *Drosophila* Bagpipe is different to XBap but not to Nkx-3.1. One of these differences is a lysine (XBap) to Arganine (Dro Bap) change. This is a conservative change and a very common variation in the homeodomains of all the NK class genes (Fig.7) and it was therefore decided not to consider this position in the first round of mutagenesis. This identifies positions 32 and 34 as potentially significant differences between XBap and *Drosophila* Bagpipe. Although these positions have not previously been implicated to have a role in specifying the DNA binding of homeodomains, the amino acids responsible for the specification of position 1 of DNA binding motifs have not yet been clearly identified.

In the XBap protein these two amino acids are alanine and leucine respectively whereas in *Drosophila* serine and methionine occupy the same positions. Analysis of the nucleotide sequence of *XBap* reveals that mutation of four nucleotides will convert the XBap amino acids to those of *Drosophila* Bagpipe at positions 32 and 34 (Figure 32.B). These mutations were introduced using the Stratagene Quick Change site directed mutagenesis kit (see materials and methods). RNA was synthesised and *in vitro* translation carried out in order to produce XMut protein. This consists of an XBap protein backbone in which the homeodomain is more like that of *Drosophila* Bagpipe, the expectation was that this mutated protein would still need two sites but the strict requirement for a 5'-TAAG-3' core over a 5'-CAAG-3' core may have been lost. The DNA binding characteristics of this protein were therefore compared with wild-type XBap protein using the panel of mutant oligonucleotides (Fig. 33).

Figure 32. Design of site directed mutation of *Xenopus Bagpipe*.

A Alignment of the amino acid sequence across homeodomains of XBap, Nkx-3.1 and Dro Bap. The two amino acids to be mutated are shown in red.

B The nucleotide sequences of XBap and Dro Bap coding for the two amino acids of interest, and the nucleotide sequence necessary to convert XBap to Dro Bap (XMut) at those positions.

C Sequence of the oligos used in the PCR based Site Directed Mutagenesis strategy.

A

```
XBap          KKRSRAAFSHAQVFELERRFNHQRYLSGPERADLAASLKLTETQVKIWFQNRRYKTKRRQ
Mus Nkx-3.1   Q.........T..I....K.S..K...A....H..KN.....................K.
Dro bagpipe   ...................A.........SEM.K..R.....................K.
```

B

XBap	A	L
	gct	ctg
DroBap	S	M
	agc	atg
XMut S	M	
	agc	atg

C

SDM Primers

GTCTGGCCCTGAGCGAAGCGACATGGCTGCCT	Top
GGAGGCAGCCATGTCGCTTCGCTCAGGGCCAG	Bottom

Figure 33 Bandshift to study the DNA binding properties of XMut protein.

A

Lanes 1-11 used T2 oligo as probe with T2, single-mutant, double-mutant, C1 and C2 oligos as competitors for binding to XMut protein.

B

Lanes 1- 5 used T2 oligo as probe with C1 and C2 oligos as competitors for binding to XBap protein as a comparison to the competition observed in A with XMut protein. Lanes 7-14 used each of the mutant oligos as probe binding to H_2O, XMut or XBap proteins.

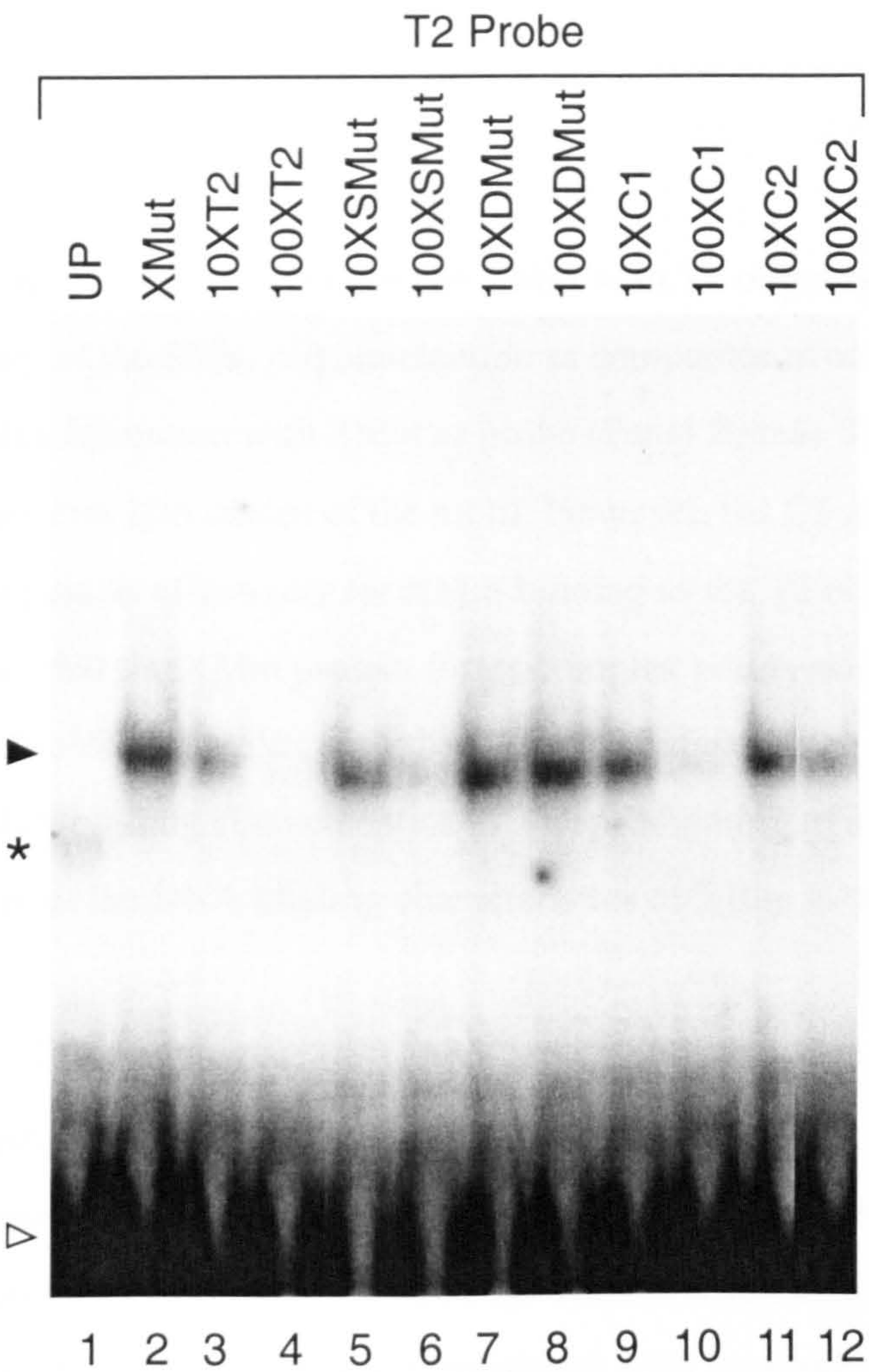
H_2O - H_2O *in vitro* translation control product.

XBap - Full length XBap protein produced by *in vitro* translation.

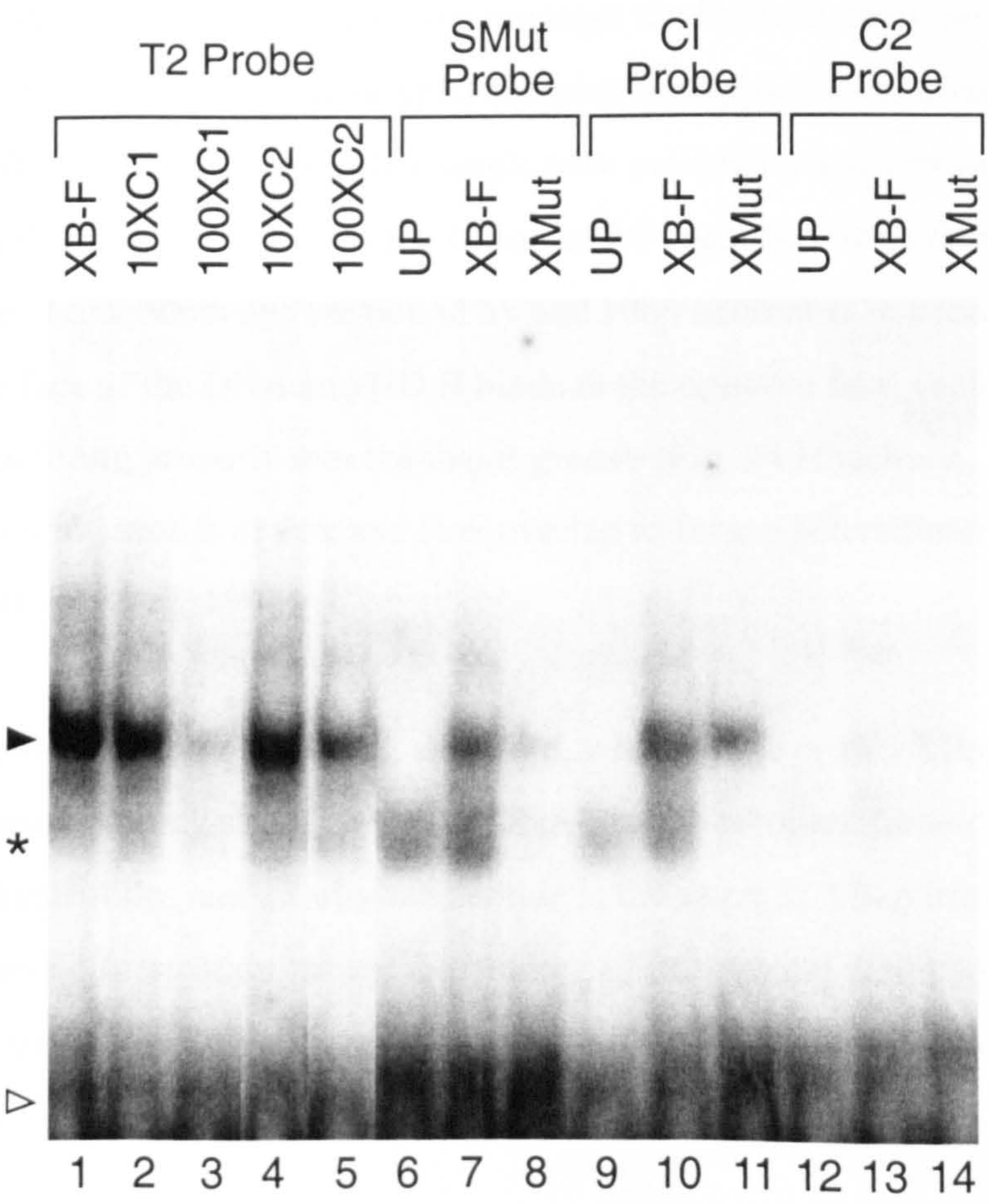
XMut - Full length XMut protein produced by *in vitro* translation.

T2	TTAAGTGG---TTAAGTGG
SMut	TTAAGTGG---TTCGATGG
DMut	TTCGATGG---TTCGATGG
C1	TTAAGTGG---TCAAGTGG
C2	TCAAGTGG---TCAGGTGG

A



B



► = specific complex
▷ = free probe
* = background

Panel A illustrates the binding of XMut with T2 oligonucleotide as probe. The reduced efficacy of the SMut oligonucleotide as competitor in comparison to T2, and reduced complex formation with SMut as probe (Panel B, lane 8) indicates that the XMut protein still requires two copies of the motif. However, the C1 and C2 oligonucleotides also fail to compete as effectively for XMut binding as the T2 oligonucleotide. This, in addition to the fact that the XMut protein forms complex poorly with the C1 probe, and fails to form any detectable complex with the C2 probe indicates that the mutated protein has retained the DNA binding characteristics of XBap. Mutation of amino acids 32 and 34 alone fails to convert the DNA binding characteristics of XBap to those of *Drosophila* Bagpipe.

5.3 Discussion

The reduced efficacy of the SMutR and SMutL oligonucleotides as competitors and their poor performance as probes clearly demonstrates that XBap requires two copies of the core motif for optimal DNA binding. The reduced binding of XBap to a palindromic arrangement of the two sites is consistent with the binding site selection data (Fig. 20). There is a precedent for such a site amongst the homeodomain proteins; the consensus DNA-binding site of even-skipped contains two approximate direct repeats AATTA and AATTC, which differ by only a single base pair (Hoey and Levine, 1988). Crystallographic studies at a resolution of 2.0 angstroms have revealed that two even-skipped homeodomains are bound by one 10bp consensus in a tandem fashion. HDI binds to one face of the DNA and HD II binds to the opposite face, each of their recognition helices fitting properly into the major groove (Fig. 34 Hirsch Aggarwal and , 1994). In addition, the two binding motifs of Pbx and Hox overlap to form a heterodimer binding site of just 12bp (Chan *et al*, 1997).

The most pressing question is, therefore, whether one or two XBap proteins bind to the consensus binding site. The size of DNA-protein complex formed with SMut oligonucleotides and T2 oligonucleotide is the same. If XBap used a system of co-operative dimerization, where the binding of one protein facilitates the binding of the other, or if two monomers bind independently of each other, we would have expected the

Figure 34. *Drosophila* Evenskipped homeodomains bind in a tandem fashion on the opposite faces of DNA

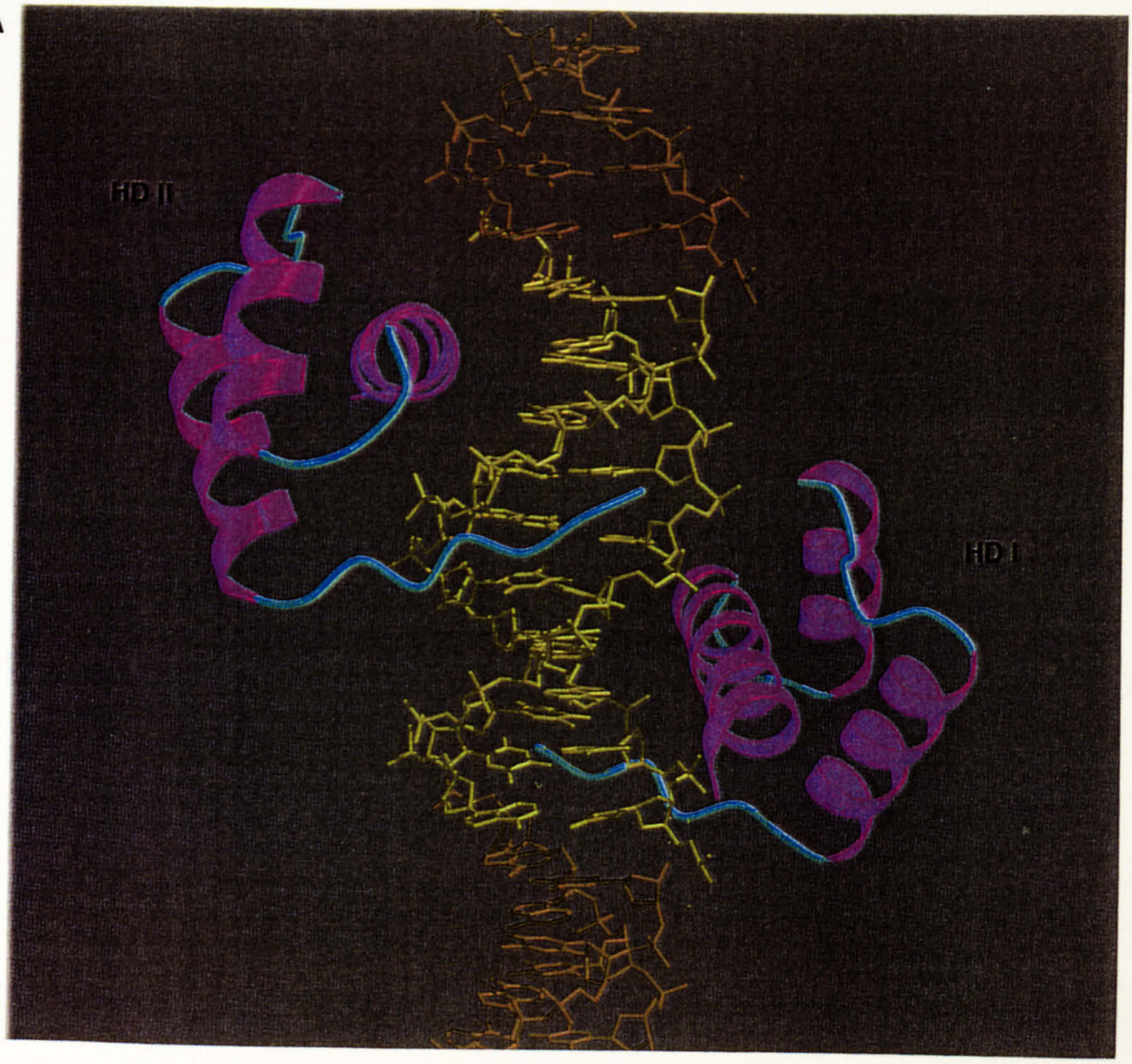
The recognition helix fits into the major groove and the flexible N-terminal arm lies in the minor groove.

A. A schematic representation of the complex. The DNA (coloured yellow) is the 10bp duplex 5'-TTAATTTAAG-3'.

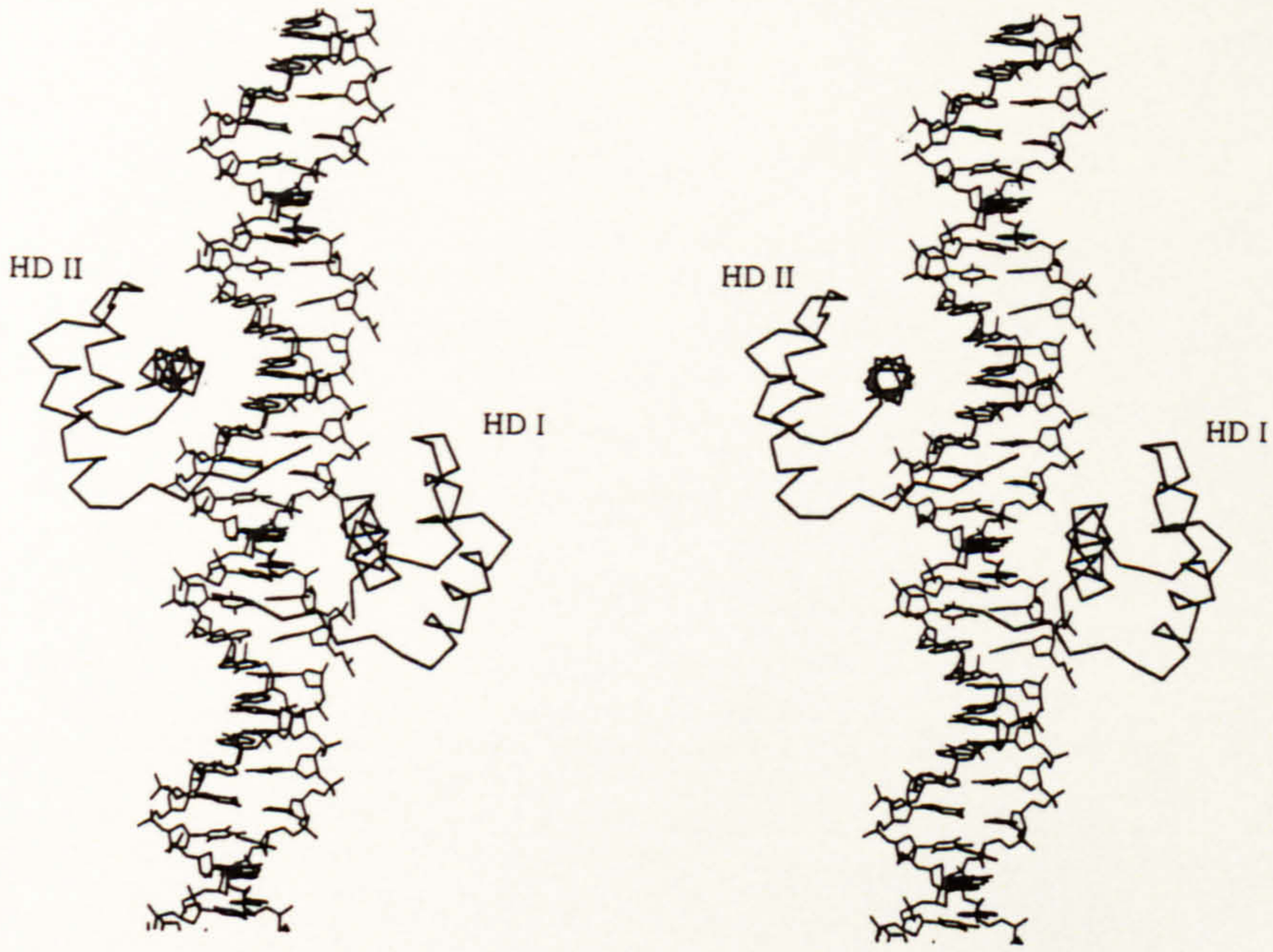
B. A stereo representation of the complex.

(Figure reproduced from Hirsch and Aggarwal, 1995)

A



B



formation of a smaller complex with the SMut oligonucleotides. This suggests that either that XBap is present in solution as a homodimer and binds to DNA as a homodimer, or that only one protein binds to the two motifs; in both cases binding affinity would be reduced when only one core motif is present. If XBap binds as a homodimer co-translation between a homeodomain only construct and the full-length construct would be expected to form three complexes; the largest corresponding to full-length homodimers, the lowest to HD homodimers and an intermediate complex corresponding to Full-HD heterodimers. Co-translation, however, fails to reveal an intermediate complex (data not shown).

This series of experiments would seem to suggest that a single XBap protein binds to the DNA binding site, although none of the experiments prove this conclusively and this would appear to be in contradiction to the precedent for this type of site. Although the formation of intermediate complexes using co-translation and EMSA has been reported to successfully identify intermediate complexes between full-length and truncated proteins (Yan *et al*, 1997 and Wilson *et al*, 1993) it is possible that full-length proteins and homeodomain proteins preferentially dimerize with each other rather than forming heterodimers. In the nematode C. elegans sequences C-terminal to the MEC-3 homeodomain are important for forming heterodimers with another homeoprotein UNC-86 and similarly, sequences C-terminal to the homeodomain in the yeast protein MATα2 are important for interactions with the a1 homeodomain (Mann , 1995). If there is such a region C-terminal of the XBap homeodomain, which enhances protein-protein interactions the full-length proteins containing this region would preferentially dimerize leaving the homeodomain proteins to dimerize with each other. Co-translation of a series of truncations may reveal a second region within the XBap protein that may contribute to protein-protein interaction. Of course, the opposite argument may also apply; the smaller size of the homeodomain alone may reduce steric constraints associated with two proteins binding such a short piece of DNA. The latter theory may also explain the fact that the homeodomain always produces DNA/protein complex more efficiently than the full-

length construct. *In vivo* there may be additional factors that act to reduce the steric constraints of the full protein or that facilitate correct folding.

The crystallographic structure of even-skipped, and the model of Ubx and Exd predicted using that study, indicate that regions within and immediately outside the homeodomain are favourably disposed for protein-protein interactions. The N-terminal arm, and the polypeptide that continues from it, are well positioned to interact with the end of helix 1, the loop between helices 1 and 2 and the end of helix 3 of the adjacent HD (Hirsch *et al*, 1995). The N-terminal region of the homeodomain is one of the patches known to be critical in the Ubx-Exd interaction (Chang *et al*, 1996), the IPOU-POU interaction and for interactions between HOX proteins (Zappavigna *et al*, 1994). Alignment of all *NK* class genes (Fig 7) shows that the *NK3* class genes are most distinguishable from their *NK2* class relatives by the sequence of their N-terminal arm. In addition amino acids 22 and 56, both implicated in protein-protein interactions (Mann *et al*, 1995) are two positions at which the NK3 class differ from the NK2 class. These differences between the NK2 and NK3 class homeodomains at positions known to be involved in protein-protein interactions could explain why only the NK3 class proteins require two DNA binding motifs for optimal binding. Co-translation of a series of truncations may reveal a second region within the XBap protein that may contribute to protein-protein interaction, and site directed mutagenesis of positions 22 and 26 to the amino acids present in NK2 class proteins may also be enlightening. There is to date no precedent for a situation in which one homeodomain binds two copies of a core motif.

Another approach to resolving this issue would be to carry out experiments to directly address whether XBap can form homodimers in solution. A common approach is to produce a fusion protein for use in an affinity chromatography column. An attempt was made to produce a XBap pGEX fusion, unfortunately, fusion protein preparations were unsuccessful. An involved but reliable assay for dimerization is the yeast-2-hybrid system (Fields and Sternglanz, 1994). This method would be most informative as in addition to

studying dimerization, additional factors that may interact with XBap in order to moderate its DNA binding specificity or biological activity may be identified.

The inability of XBap to form DNA/protein complex with the C2 oligonucleotide, in addition to the reduced efficacy of C1 and C2 as competitors for T2 binding, clearly demonstrates that XBap requires two copies of the 5'-TAAG-3' core for optimal DNA binding. Bapx1, which is identical across the homeodomain with XBap has the same DNA binding requirements as XBap; two copies of a 5'-TAAG-3' core. Nkx-3.1, which has some differences across the homeodomain compared to XBap (Fig. 14), also has the same requirements. *Drosophila* Bagpipe, however, is unusual in that although it requires two core motifs it can form a DNA/protein complex with an oligonucleotide containing 5'-CAAG-3'cores, although at a reduced affinity compared to 5'-TAAG-3'.

Sciavolino *et al*, 1997, have studied the DNA binding of Nkx-3.1, a murine NK3 class protein, by EMSA to a panel of oligonucleotides. This group demonstrated that Nkx-3.1 interacts preferentially with sites containing a 5'-CAAG-3' core over a 5'-TAAT-3' core; this observation may suggest that the G at position 4 is more discriminating than the T at position 1 in the binding of NK3 class proteins. This would be consistent with the increased affinity of TTF-1 for an Antp site mutant in which only one position of the wild-type Antp site had been mutated to produce a 5'-TAAG-3' core motif, the increased affinity being greater than that observed with a 5'-CAAT-3' motif in the same context (Damante *et al*, 1994). Although binding of Nkx-3.1 to 5'-CAAG-3'may appear to contradict the data presented here, the group did not test a site containing 5'-TAAG-3' against which to compare binding of the 5'-CAAG-3' core.

Although TTF-1 has been reported to discriminate between specific and non specific DNA binding motifs more efficiently under salt and temperature conditions close to an intracellular environment (Damante *et al*, 1993) the DNA binding preferences of XBap and *Drosophila* Bagpipe were unaltered under a range of salt conditions. *Drosophila* Bagpipe was able to form DNA/protein complex with the C2 oligonucleotide under all of

the conditions tested. The observed difference in the binding of XBap and *Drosophila* Bagpipe is therefore not likely to be an artefact of the salt conditions.

It has been demonstrated that XBap DNA binding specificity is dictated by the homeodomain and, therefore, the amino acids responsible for the difference in binding between *Drosophila* Bagpipe and XBap are most likely to be in this domain. An alignment of XBap, Nkx-3.1 and *Drosophila* Bagpipe revealed two amino acids that seemed to represent the greatest deviation from the XBap and Nkx-3.1 amino acid sequences (Fig.32) but site directed mutagenesis of these key amino acids failed to alter the preference of XBap for a 5'-TAAG-3'core.

Although previous studies have identified a single amino acid responsible for the specification of base pair 4 of homeodomain DNA binding sites (Damante *et al*, 1996), specification of base pair 1 is not so clear. Amino acids at positions 3 and 5 of the homeodomain N-terminal arm have been reported by several groups to interact via the minor groove with base pairs at positions 1 and 2 of a 5'-$Y_1A_2A_3K_4$-3' DNA binding site (Wilson *et al*, 1996, Mann *et al*, 1995 and references therein). However, the situation is not as simple as proposed by Wilson *et al*, who have suggested a 5'-TAAT-3' binding consensus. According to this consensus basic amino acids at positions 2 and 3 of the HD and an Arginine residue at position 5 confers binding to a T at position 1 (Wilson *et al*, 1996). However, this is obviously an oversimplification as TTF-1, XBap and *Drosophila* Bagpipe all contain this sequence of amino acids but their binding to position 1 is variable.

Damante *et al* had previously demonstrated that residues 1-28 of the TTF-1 HD contribute to DNA binding specificity (Damante and Di Lauro, 1991). This group then carried out a study of several TTF-1 HD mutants in which amino acids of this region were changed to those of Antp. A change in binding properties was observed with the mutant TTF-1 HD(QTY) in which Val6, Leu7 and Phe8 were changed to Gln, Thr and Tyr respectively (these amino acids are present at position 6, 7 and 8 in AntpHD). TTF-

1HD(QTY), in contrast to the wild type TTF-1 HD, prefers a 5'-TAAG-3' DNA binding sequence, although it can still bind efficiently to a 5'-CAAG-3' core (47% of DNA binding activity of WT TTF-1 to its consensus binding site) (Damante *et al*,1996). XBap and *Drosophila* Bagpipe both have the sequence Ala, Ala, Phe at positions 6, 7 and 8; a sequence more closely related to the Val, Leu, Phe of TTF-1 than the Gln, Thr, Trp of Antp. It is surprising in the light of Damante *et al*'s study, that XBap should prefer a T at position 1 of its DNA binding sites rather than a C. There are obviously additional amino acids in the homeodomain that contribute to the specification of this position of DNA binding.

The crystallographic resolution of the even-skipped protein complexed with DNA revealed that the N-terminal arm grasps the DNA in the minor groove (Hirsch *et al*, 1995). The trajectory of the N-terminal arm differs from that in the Engrailed (En) and Antennapedia (Antp) structures. It is proposed that this difference may be due to Tyr4 and its interaction with the adenine of bp 3. Tyr4 is unique to the Eve class of homeoproteins (En contains a proline and Antp contains a glycine). Due to variability in the N-terminal paths, the highly conserved residue Arg3 interacts with DNA differently in the three structures. Judging from the structure elucidated by Hirsch *et al*, 1995, the group propose that the identity of residue 4 may be particularly important in determining the path of the N-terminal arm and the consequent DNA binding differences. This is particularly significant when studying the differences between NK2 and NK3 proteins; in the NK3 class residue four is a highly conserved serine but in the NK2s residue 4 is either a Proline or Arganine. This suggests that the trajectory of the N-terminal arm may be different between NK3 and NK2 genes resulting in differences in DNA-protein interactions. Considering that the N-terminal arm has also been implicated in mediating protein-protein interactions (Chang *et al*, 1996, Zappavigna *et al*, 1994) it would be interesting to observe the effect of an N-terminal arm swap on the binding properties of NK2 and NK3 proteins.

5.4 Future experiments

Pellizari *et al*, have noted that an amino acid that contributes to the function of a protein in one context may be unfavourable in another (study comparing amino acids at positions 50 and 54, Pellizari *et al*, 1997). In other words, the role of an amino acid at a particular position may depend on amino acids at other positions within that DNA binding region. The assumption that the positions at which both Nkx-3.1 and *Drosophila* Bagpipe differ from XBap do not contribute to specificity may, therefore, be flawed. In addition, the original study by Damante had implicated amino acids in two separate regions (1-28 and 50-60) to be responsible for position 1 binding specificity. An analysis of the alignment of XBap with the NK-2 class proteins that bind 5'-CAAG-3' reveals that 9 out of the 12 differences between the two lie within these regions.

The next step would be to construct chimaeric proteins with varying proportions of XBap and TTF-1 homeodomain sequences, starting with an N-terminal arm swap, in order to identify regions, or combinations of regions, responsible for the difference in specificity of the two proteins. Specific amino acids in these regions could then be analysed more closely using site-directed-mutagenesis.

Having determined that XBap can bind to DNA, and defining its DNA binding site, an assay needed to be conducted that would determine if the binding of XBap protein to DNA could have an effect on the transcription of downstream genes.

Chapter 6

Study of the transcriptional abilities of XBap

6.0 Chapter 6: Study of the transcriptional abilities of XBap

6.1 Introduction

Homeodomain containing transcription factors control various cell fate decisions in a wide range of organisms including yeast, insects and vertebrates. XBap, therefore, may function in the differentiation of visceral mesoderm either by activating transcription of specific genes or by repressing transcription of genes not compatible with a visceral mesoderm cell fate.

Several NK-2 class proteins had previously been identified as potential transcriptional activators by cell transfection with reporters containing multimerized Nkx-2.5 DNA binding sites (Reecy *et al*, 1997, Chen *et al*, 1995). It was, therefore, decided to take advantage of the *Xenopus* oocyte assay in order to investigate whether XBap can also act as a transcriptional activator.

Previous studies had shown the optimal DNA binding site of rat TTF-1 to be 5'-TNNAAGTG-3' (Guazzi *et al,* 1990) and *in vitro* binding site selection studies had yielded a similar consensus site, 5'-CAAG-3', for Nkx-2.5, XNkx-2.3 and 2.5 (Chen *et al*, 1995 and Mohun *et al,* unpublished data). Chen *et al* had also demonstrated binding of Nkx-2.5 to the common homeodomain binding site core 5'-TAAT-3' (referred to as A20). At the beginning of this study the DNA binding sites of *Drosophila* Bagpipe and XBap had not yet been identified. It was therefore decided to first determine whether XBap could activate transcription from either 5'-CAAG-3' or 5'-TAAT-3' sites in a *Xenopus* oocyte assay, with the optimal XBap DNA binding sites being studied once they had been identified.

6.2 Results

6.2.1 TTF-1 binding site *Xenopus* oocyte assay.

Full length XBap RNA was synthesised from ORF/Myc/pT7 and 5-10ng injected into the vegetal pole of oocytes. The oocytes were incubated at 18°C for at least 12 hours to allow

translation of the RNA. 5ng of pBLCAT3T reporter DNA containing 1 or 5 TTF-1 sites, 5 mutated TTF-1 sites, and 1 or 4 A20 sites was then injected into the germinal vesicle (Fig. 8). In each assay approximately 100 oocytes were injected with each RNA/reporter combination, each group being divided into three before the protein extract was prepared from the oocytes and assayed for CAT activity. Three assays were carried out using these reporters and figure 35 illustrates a representative result.

In all the experiments, constructs containing concatemerized sites were more effective than those with a single binding site (compare lanes 6 and 8). Activation levels induced by XBap with TTF-1 sites are only 3-4 fold above background levels (lane 8 compared to lane 2) and not above background at all with A20 sites (compare lane 12 and lane 2). It is questionable whether such a low activation level is biologically significant. This result, however, was consistent with the weak activation observed with XNkx-2.3 and XNkx-2.5 from the same reporters (Mohun *et al* unpublished observations).

6.2.2 XBap binding site *Xenopus* oocyte assay.

A study carried out by Chen *et al*, 1995, had shown that full length murine Nkx-2.5 can serve as a modest transcriptional activator in transfected 10T1/2 fibroblasts. However, deletion of the amino acids C-terminal of the homeodomain appeared to increase the transcriptional activation of the reporter gene by approximately 20 fold (Chen *et al*, 1995). A reporter containing three copies of the recently defined XBap DNA binding site (3XBpBLCAT3T) was, therefore, tested in a *Xenopus* oocyte assay with three constructs; full length XBap protein (ORF/Myc/pT7), the N-terminal portion and the homeodomain of XBap (C-Del/pT7) and the homeodomain and the C-terminal portion of XBap (N-Del/pSP). SL1 is a *Xenopus* member of the MEF2 family of transcription factors, which can activate transcription of genes containing CANNTG binding sites. Injection of SL1 and pBLCAT3T reporter containing three copies of is binding site were used as a positive control of oocyte quality with every assay and always gave very strong activation (data not shown)(Chambers *et al*, 1994).

Figure 35. A *Xenopus* oocyte transactivation assay using a pBLCAT3T reporter containing TTF-1 binding sites.

y axis - % conversion of chloramphenicol to acetylated chloramphenicol.

x axis - DNA/RNA combinations injected in each sample group (5-10ng RNA, 5ng DNA).

0T - pBLCAT3T with no TTF-1 DNA binding sites

1T - pBLCAT3T with 1 TTF-1 DNA binding site

5T - pBLCAT3T with 5 TTF-1 DNA binding sites

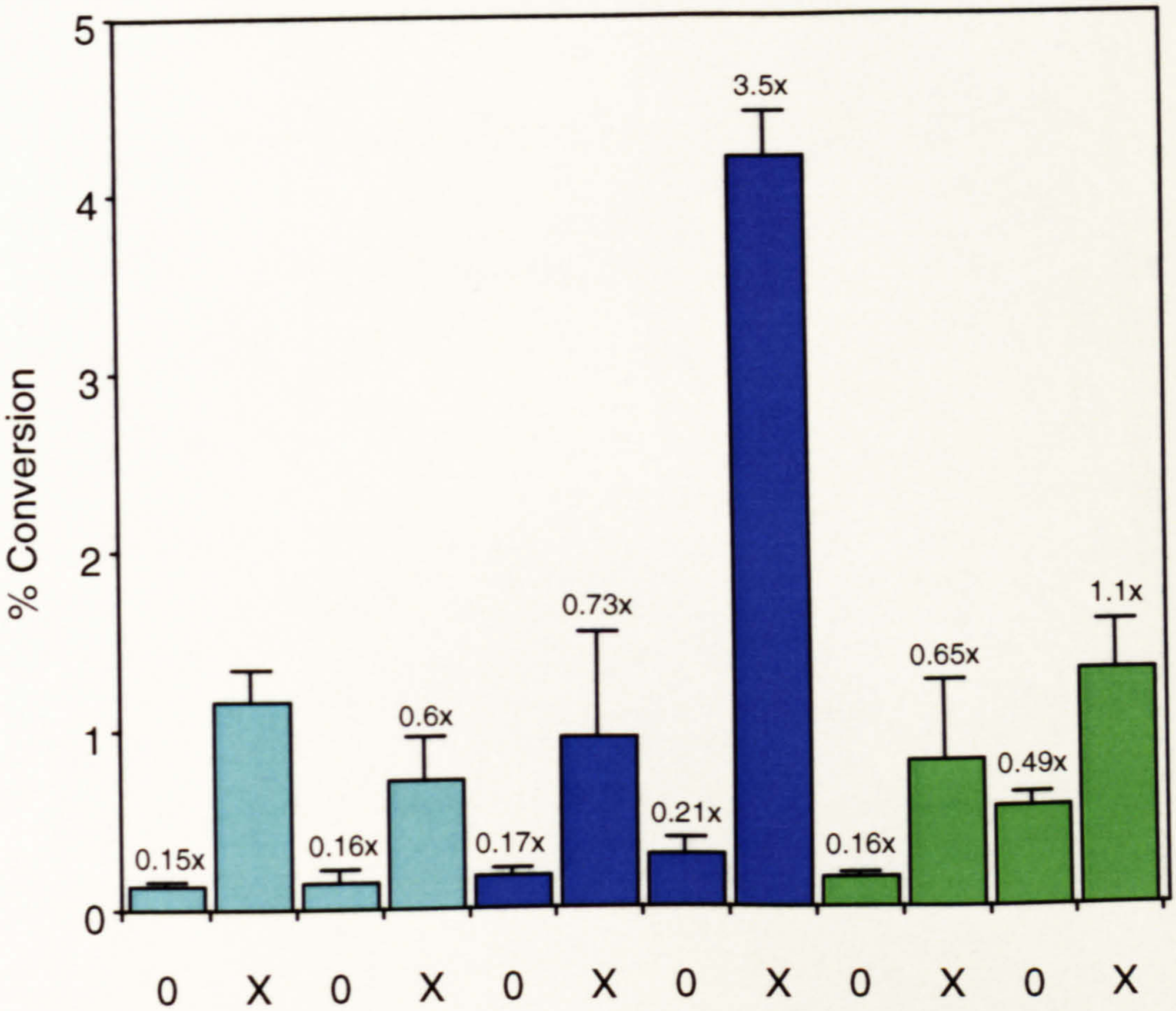
1A - pBLCAT3T with 1 A20 DNA binding site

4A - pBLCAT3T with 4 A20 DNA binding sites

MutT - pBLCAT3T with 5 mutated TTf-1 binding sites

X - full length XBap synthetic RNA produced from the ORF/Myc/pT7 plasmid.

Error bars - Standard Deviation.


% Conversion
5
4
3
2
1
0
3.5x
0.15x
0.16x
0.6x
0.17x
0.73x
0.21x
0.16x
0.65x
0.49x
1.1x
RNA
0 X 0 X 0 X 0 X 0 X 0 X
Binding sites in reporter
0T 0T MutT MutT 1T 1T 5T 5T 1A 1A 4A 4A

A total of six assays were carried out with full length XBap, three with the N-Del construct and three with the C-Del construct. The results of the six assays were so variable that no single experiment is representative, therefore, the results of all the assays are illustrated in tabular form in figure 36. Activation by the full length construct varies between 0 and 8.3 fold, the N-Del construct between 0 and 2 fold and C-Del did not activate at all. Due to the lack of consistency between assays it is impossible to make any firm conclusion as to the transcriptional activation abilities.

XBap protein of the correct size had been produced *in vitro* from the ORF/Myc/pT7 construct, however, in order to ascertain that the construct was being translated correctly *in vivo* a western blot of injected oocyte extract was carried out (Fig. 37). Although the washes of this blot were not as stringent as they should have been this assay was sufficient to demonstrate that the construct was being correctly translated in *Xenopus* oocytes.

6.2.3 3T3-Fibroblast transfection with pBLCAT3T reporter

It was possible that the failure to detect activation by XBap could be an effect specific to the oocyte assay system. Transfecting XBap into NIH 3T3 fibroblasts addressed this possibility. A β-galactosidase reporter was used to normalise against the efficiency of transfection between the samples. An Nkx-2.5/VP16 fusion which had been shown to strongly activate transcription from the TTF-1/pBLCAT3T reporter in the oocyte system was used as a positive control of both the reporter construct in this system and of the extraction procedure. Activation of the 3XBapbinding sites/pBLCAT3T reporter was detected in cells transfected with the Nkx-2.5/VP16 fusion but full length XBap (ORF/Myc/pC) failed to show any activation of this reporter in transfected cells (data not shown).

It was possible that although the construct was being translated correctly, it was not being localised to the nucleus. In order to address this issue full length XBap protein tagged with the myc epitope (ORF/Myc/pC) and a homeodomain only construct tagged with the

Figure 36. Table of results of six *Xenopus* oocyte assays using a pBLCAT3T reporter containing XBap DNA binding sites with three different *XBap* synthetic RNAs.

Full - synthetic RNA produced from ORF/Myc/pT7.

C-Del - synthetic RNA produced from C-Del/pT7.

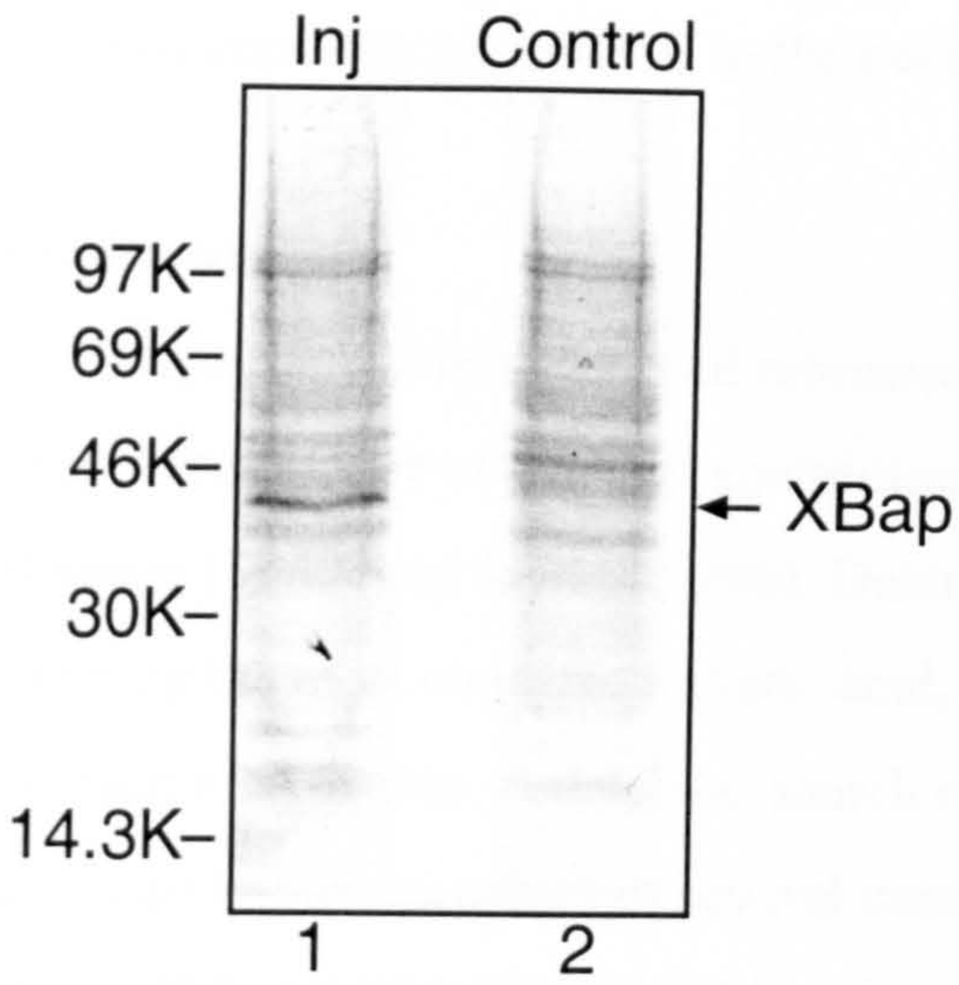
N-Del - synthetic RNA N-Del/pSP.

No of XBap DNA Binding sites in reporter	0	3	0	0	0	3	3	3			
RNA injected	0	0	Full	C-Del	N-Del	Full	C-Del	N-Del	Full	C-Del	N-Del
	% Conversion of chloramphenicol to acetylated chloramphenicol								Fold Increase		
Assay A	0.27	2.18	0.65	0.55		2.79	1.10		1.28	0	
Assay B	1.02	1.05	0.82	0.48		1.02	0.42		0	0	
Assay C	0.16	0.63	0.48	0.60		0.98	0.45		0	0	
Assay D	0.80	1.30	0.67		0.90	10.89		2.16	8.3		1.66
Assay E	1.49	1.48	1.15		0.90	5.70		3.17	3.8		2.1
Assay F	1.09	1.86	0.68		1.09	1.02		2.29	0		1.23

Figure 37. Western blot of *XBap*-injected and uninjected oocyte extracts.

Lane 1 Western blot of extract from oocytes injected with RNA produced from DNA construct ORF/Myc/pT7.

Lane 2 Western blot of extract from uninjected oocytes taken from the same batch of oocytes as those used for injection.


Inj
Control
97K–
69K–
46K–
30K–
14.3K–
XBap
1
2

Ha epitope (HD/Ha/pC) were transfected into 3T3 fibroblasts. These cells were then fixed in paraformaldehyde followed by immunocytochemical processing with 9E10 or anti-Ha as primary antibody and Flourescin (FITC) conjugated affinity pure goat anti-mouse IgG as secondary antibody. Cells transfected with pCDNA3 alone were processed with Myc and Ha antibody in order to measure the background fluorescence due to each antibody. Figure 38 demonstrates that although there is protein present in the cytoplasm, both the full length and homeodomain constructs are localised to the nucleus. This result suggests that the nuclear localisation signal is within the homeodomain and, therefore, that both of the deleted constructs should also be correctly transported to the nucleus.

activation or

6.2.4 Transcriptional repression assay

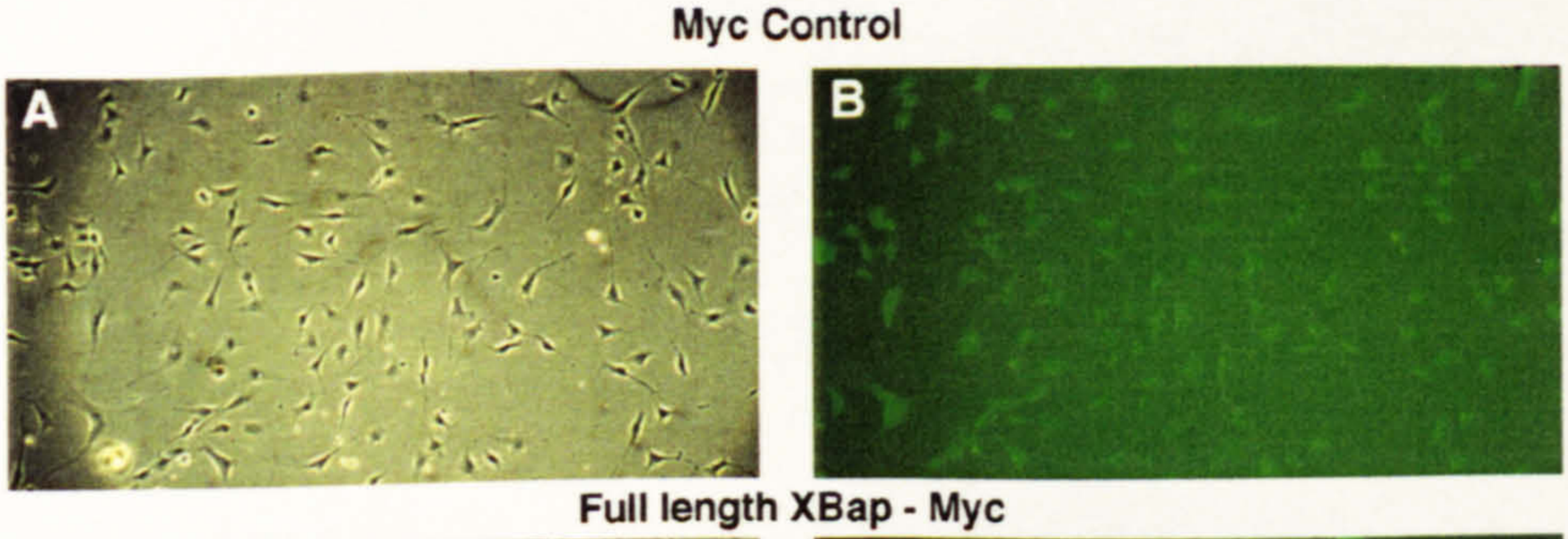
A study carried out by Smith and Jaynes into transcriptional repression by *Drosophila* Engrailed demonstrated the critical importance of a region containing the only conserved domain outside of the homeodomain (Smith and Jaynes, 1996). Deleting either the core of this homology region, or mutating the most conserved amino acid, Phe 175, strongly decreases repressive activity of Engrailed *in vivo*. A database search revealed regions homologous to this eh1 domain in all known members of several classes of homeobox containing genes including *NK-1*, *NK2* and *NK3* (Smith and Jaynes, 1996). The eh1 domain is homologous to what has been known as the 'TN' domain in the N-terminal region of NK class genes. An alignment of NK3 class sequences in the region of the TN domain demonstrates a significant conservation with the consensus eh1 sequence produced by Smith and Jaynes (Fig. 39). Additionally, a recent study of *Drosophila clift* revealed that the dorsoventral position of SGP specification is refined by negative interaction with the dorsally adjacent visceral mesoderm precursors, and that wild-type *bagpipe* may be responsible for this refinement. Although there is no evidence that *bagpipe* has a direct effect on *clift* expression, this study and the presence of the eh-1 like domain suggest that on binding to regulatory regions XBap might in fact have a repressive effect on transcription.

Figure 38. Immunocytochemistry of 3T3-fibroblasts cells transfected with full length XBap tagged with Myc epitope or XBap homeodomain tagged with Ha.

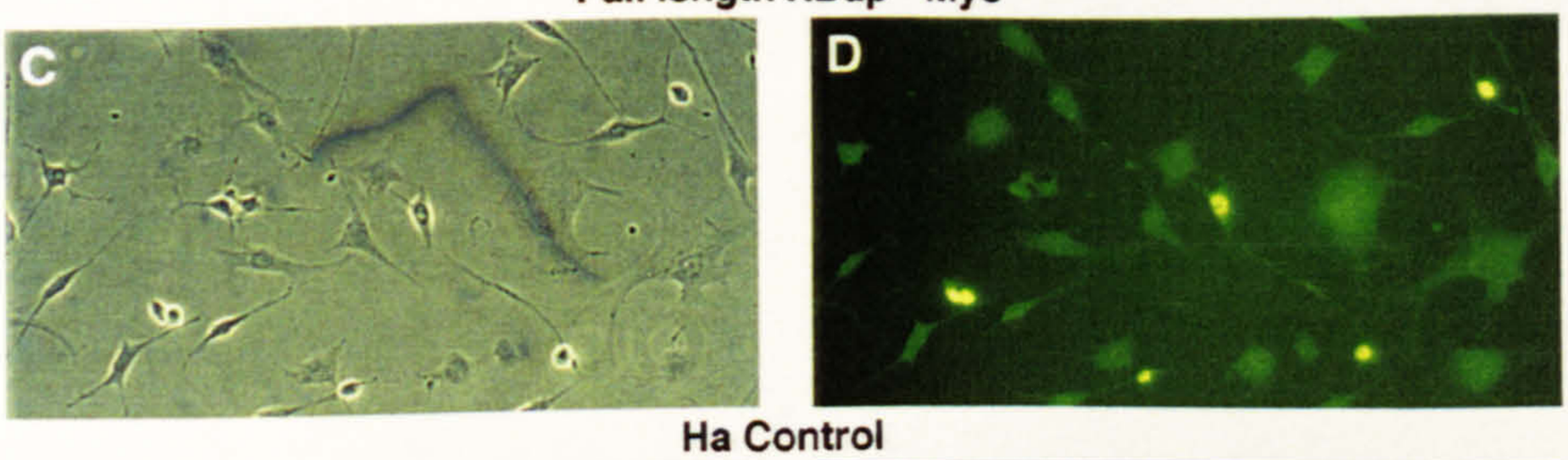
Myc control and Ha control were transfected with pCDNA3 and treated in parallel with the XBap transfections with either anti-myc or anti-Ha antibody. Full length XBap was transfected using the ORF/Myc/pC construct and XBap homeodomain with the HD/Ha/pC construct.

Photographs on the left were taken under bright field and the right with fluorescence. The controls were photographed at 10, 000X magnification and the XBap transfections at 20,000Xmagnification.

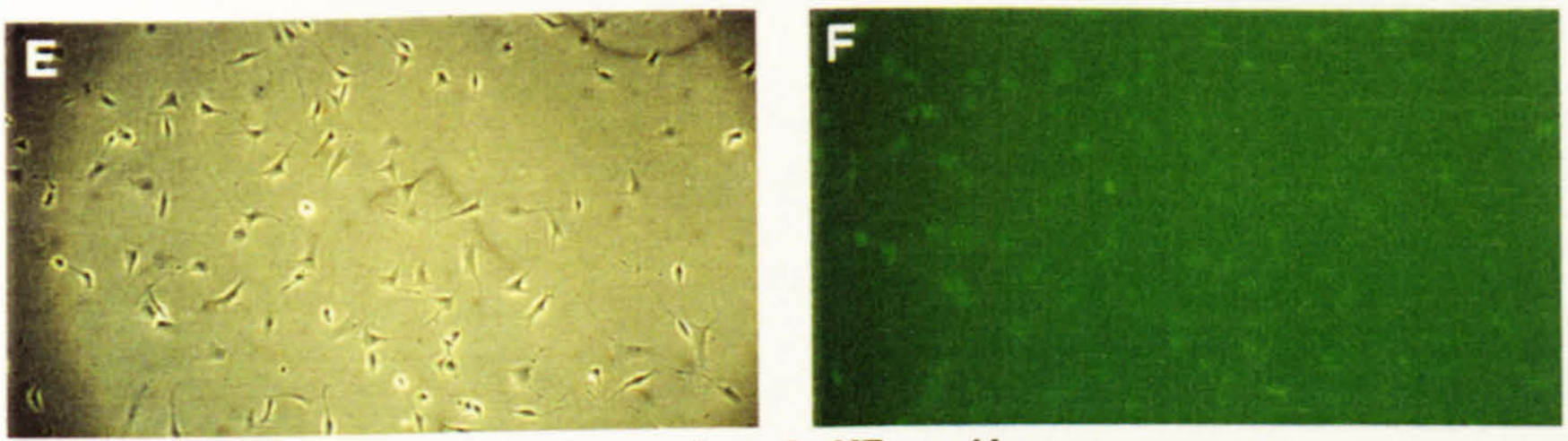
Myc Control



Full length XBap - Myc



Ha Control



Homeodomain XBap - Ha

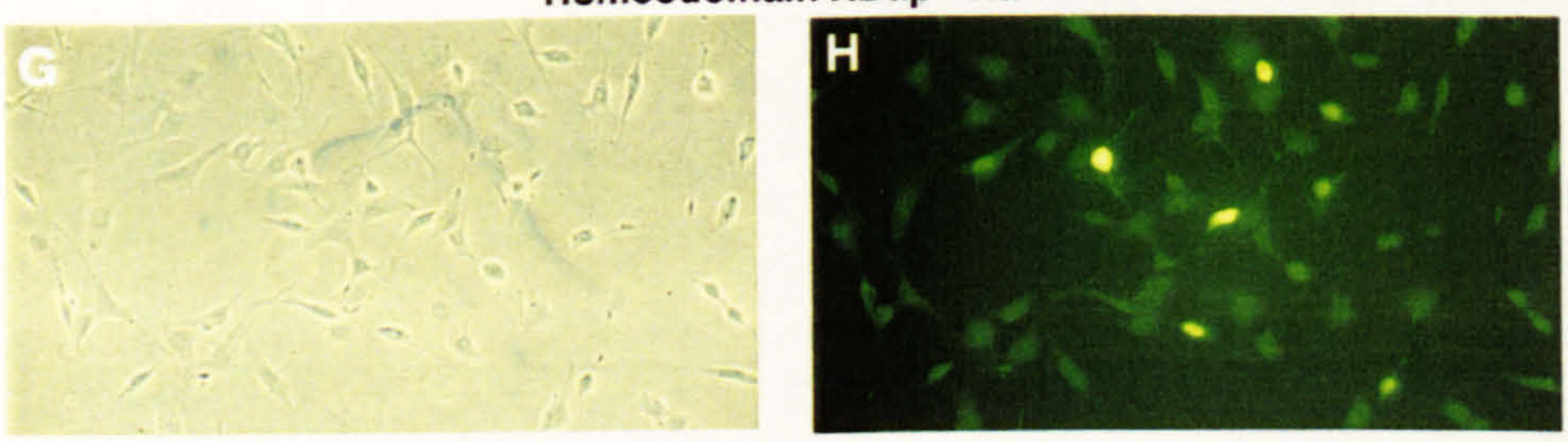
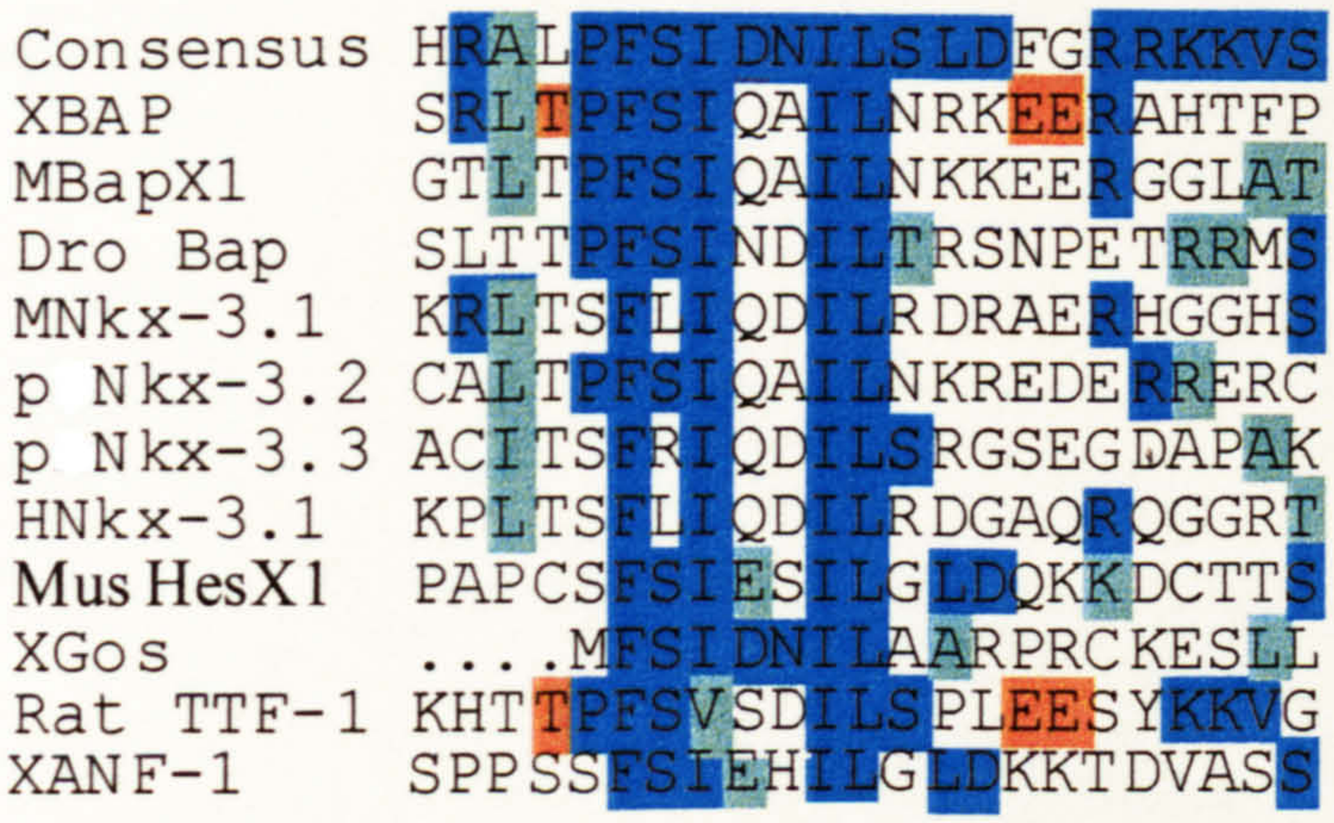


Figure 39. Comparison of the putative inhibitory domains of NK3 class genes with consensus inhibitory domain derived from studies of engrailed, Hesx1 and goosecoid

Gene names are prefixed with a species abbreviation: m, mouse: Dro, *Drosophila*: pn, *planaria*: p, *pleurodeles*: X, *Xenopus*: h, human.

Murine Bapx1 and human bapx1, Triboli et al, 1997, murine Nkx-3.1, Bieberich et al, 1996, human Nkx-3.1, Prescott et al, 1998, *planaria* bagpipe, Balavoine et al, 1996, *pleurodeles* Nkx-3.2, and Nkx-3.3, Nicolas et al, 1999, NK-3, Kim et al, 1989, XANF-1, Hesx1 and Goosecoid, Smith and Jaynes 1996, and references therin, rat ttf-1, Guazzi et al, 1990.


Consensus HRALPFSIDNILSLDFGRRKKVS
XBAP SRLTPFSIQAILNRKEERAHTFP
MBapX1 GTLTPFSIQAILNKKEERGGLAT
Dro Bap SLTTPFSINDILTRSNPETRRMS
MNkx-3.1 KRLTSFLIQDILRDRAERHGGHS
p Nkx-3.2 CALTPFSIQAILNKREDERRERC
p Nkx-3.3 ACITSFRIQDILSRGSEGDAPAK
HNkx-3.1 KPLTSFLIQDILRDGAQRQGGRT
Mus HesX1 PAPCSFSIESILGLDQKKDCTTS
XGos ....MFSIDNILAARPRCKESLL
Rat TTF-1 KHTTPFSVSDILSPLEESYKKVG
XANF-1 SPPSSFSIEHILGLDKKTDVASS

Preliminary studies were carried out to address this possibility using a cell transfection system in which an SV40 early promoter drives a firefly luciferase reporter. On transfection into ES cells this reporter causes a level of activation that may be either increased or decreased by co-transfection of Gal4DNA binding domain (GAL4DNB) fusion's by virtue of the GAL4 DNA binding sites upstream of the SV40 promoter (vectors unpublished, kindly supplied by Dr. Joshua Brickman, NIMR). GAL4DNB fusion's of full-length XBap and of XBap/CDel were constructed.

Each transfection was carried out in duplicate with 100ng of the Firefly luciferase/GAL4DNB/SV40 reporter and 100ng of pRL-SV40 reporter construct (Promega) against which to normalise transfection efficiency. 100ng, 200ng and 400ng of each GAL4DNB fusion construct was transfected. Gal4DNB alone was unable to activate transcription of the reporter and was used to set the reference level of firefly luciferase activity against which the GAL4DNB fusion's are compared. Residues 1-85 of the murine HesX1 (including eh1 domain) fused to GAL4DNB provided a positive control of repression. Transfection of residues 50-186 (eh1 domain deleted) of HesX1 fused to GAL4DNB demonstrated that the repression of the reporter is specific and not due to a non-specific effect that could be caused by any GAL4DNB fusion protein.

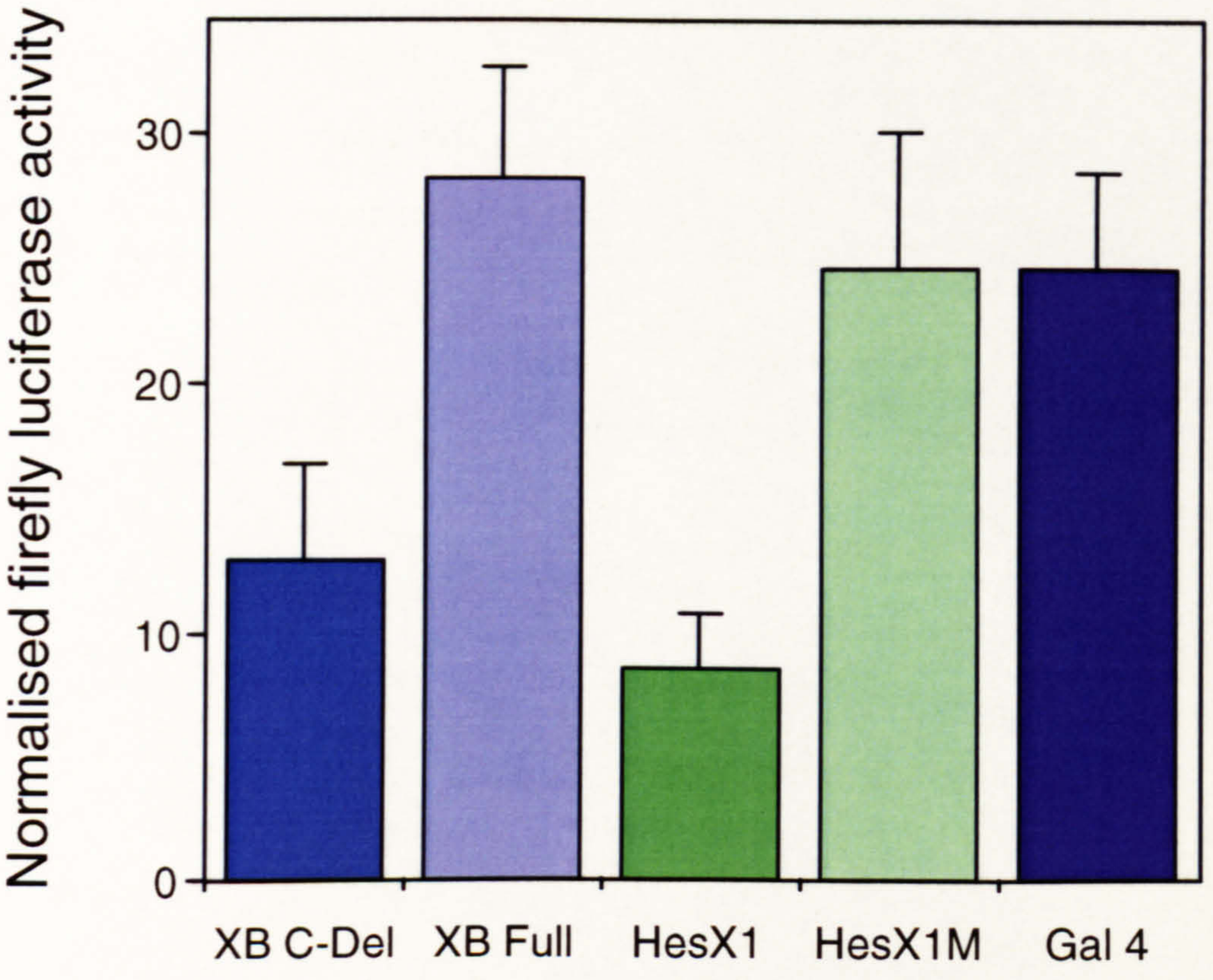
The results are illustrated in figure 40. Each column represents the average of the duplicates of 3 DNA levels as there was no variation over 100-400ng DNA amounts. Co-transfection of a full length XBap/Gal4DNB fusion with this reporter has little effect on activation (Column 2). However, transfection of a construct containing only the homeodomain and the N-Terminus containing the 'TN 'domain (XBap C-Del), causes transcription of the reporter to be significantly reduced. Each of these constructs has been tested twice and the same effect observed. Therefore, under the conditions of this assay, a Gal4DNB fusion of a C-terminal deletion of XBap acts as a transcriptional repressor.

Figure 40. The effect of Gal-4 fusion constructs on the activity of a reporter construct containing GAL4 DNA binding sites upstream of an SV40 promoter controlling firefly luciferase, in transfected ES cells.

x axis Gal4 fusion transfected.

y axis Firefly luciferase activity normalised against cell transfection efficiency with Renilla luciferase activity.

Error bars - Standard Deviation.


Normalised firefly luciferase activity
30
20
10
0
XB C-Del
XB Full
HesX1
HesX1M
Gal 4
Gal4 construct

6.3 Discussion

NK class proteins had been demonstrated to be transcriptional activators by several groups (Chen *et al*, 1995, Reecy *et al*, 1997, Mohun *et al*, unpublished data). However the *Xenopus* oocyte assay used to test XBap with TTF-1 DNA binding sites and XBap binding sites revealed very low activation with both sites and extremely variable results with the XBap sites. A western blot demonstrated that the construct was correctly translated *in vivo*, 3T3 fibroblast transfections showed that the lack of activation was not oocyte specific and immunocytochemistry showed that, in mammalian cells at least, the constructs were correctly localised to the nucleus. Testing of XBap in an assay which can detect either transcriptional activation or repression by transfected proteins however, revealed that a full-length XBap/GAL4DNB fusion has no effect on transcription but a GAL4DNB fusion containing only the N-terminus and the homeodomain acts as a repressor of transcription in this system. In this assay the removal of the C-terminus appears to reveal a domain with potential transcriptional repression activity in the N-terminus.

The NK-2 class protein TTF-1 has a TN domain homologous to that of XBap and to the eh-1 domain (Fig.39) and studies of its transcriptional activity may set a precedent for the presence of both transcriptional activation and repressor domains within a larger NK class protein. De Felice *et al*, 1995, carried out an extensive study to map the regions of TTF-1 involved in transcriptional activation in which expression vectors containing a series of TTF-1 deletions were transfected into HeLa cells with a reporter containing the Thyroglobulin (Tg) promoter upstream of CAT. These deletions indicated the presence of a transcriptional activation domain N-terminal to the homeodomain, an essential component of which appeared to be between amino acids 1 and 50 (the region containing the TN/eh1 domain). However, when these same deletions were tested with the artificial C5 E1b promoter, in which 5 TTF-1 binding sites are arranged head to tail upstream of a the E1b TATA box, all of the mutants behaved in the same manner except the deletion of AA1-50. In the context of the C5 E1b promoter the AA1-50 deletion produced a transcriptional activity higher than that of wild-type protein (De Felice *et al*, 1995).

In order to obtain an independent assessment of the functional domains in TTF-1, constructs encoding chimaeric proteins containing different segments of TTF-1 fused to the GAL-4 DNA binding domain were prepared (De Felice *et al*, 1995). TTF-1/GAL4DNB fusion's were tested on the G5E1b reporter construct containing 5 GAL4 DNA binding sites upstream of the E1b TATA box. Consistent with the Thyroglobulin promoter results, the region of TTF-1 N-terminal to the HD contained an autonomous transactivating domain, however, the limits of this domain were located at the N-terminal side between amino acids 51 and 102 and at the COOH terminus between amino acids 90 and 123. In agreement with the results obtained with the deletion mutants on the C5 E1b promoter, removal of the region spanning AA1-51 determined an increase in transcriptional activity. Thus, in two different contexts, region 1-51 displays an inhibitory effect on TTF-1 transcriptional activation. However, in FRTL-5 cells (differentiated thyroid cells) this domain appears to have an activating function. These results suggest that this region of TTF-1 is capable of behaving in both a cell-type (HeLa versus FRTL-5) and promoter (Thyroglobulin versus C5E1b) dependent manner (De Felice *et al*, 1995).

A second example in which the eh-1 domain appears to be required for only a subset of the whole proteins function is *Drosophila* Goosecoid in which the GEH/eh-1 domain is necessary for rescue of UV-ventralised *Xenopus* embryos but is dispensable for ectopic expression of Xlim-1 (Mailhos *et al*, 1998). These two examples suggest that the TN/eh1 domain can act as a transcriptional repressor as part of a larger protein, which may also have transcriptional activation abilities.

Study of *Drosophila* Goosecoid (D-Gsc) has provided evidence that eh1 containing proteins may be able to repress transcription using a mechanism involving protein-protein interactions (Mailhos *et al*, 1998). D-Gsc is a member of the paired class of homeoproteins which can co-operatively dimerize on palindromic binding sites containing two core 5'-TAAT-3' HD sites (P3). D-Gsc can achieve passive repression by

competing for the same binding sites as other K50 homeoproteins (as in the repression of Bicoid activation) or by active repression by binding to neighbouring binding sites (as in D-Gsc repression of GR activation). Mailhos *et al* carried out a GST pull-down assay between Orthodenticle (a paired-class homeoprotein) and either wild-type D-Gsc, or a D-Gsc mutant in which the GEH (Goosecoid Engrailed Homology) domain has been deleted. The study revealed that the ability of D-Gsc to strongly repress the activation of GR by Orthodenticle relies on specific protein-protein interactions for which the GEH domain and the N-terminal arm/helix 1 region of the homeodomain are both required. The group proposes a mechanism of 'interactive repression' in which repression is mediated by protein-protein interactions between repressor and activator homeoproteins (Mailhos *et al*, 1998).

In each class for which several members are known, a distinct version of the eh1 motif is recognisable, for example, within the classes expressed predominantly in mesodermal derivatives. In NK2 a Val is found at the 2nd position after the invariant Phe followed by a basic residue at the next position whereas most other classes contain an acidic residue at this position (Smith *et al*, 1996). NK3 class proteins are different to both NK2 and the consensus, containing a glutamine at this position (3rd after the Phe). Such patterns of conservation have been suggested by Smith *et al* to reflect independently conserved functional divergence among the classes, such as might occur in conjunction with the evolution of a family of interacting proteins. This theory allows that there may be significant divergence of function associated with each distinct partner in the putative interacting family. It may be that as the interacting partner of each eh-1 class protein diverged, the transcriptional activity was modified and perhaps reversed (Smith *et al*, 1996). This may explain the apparent disparity that the eh1 containing Nkx-2.5 can function as an activator of transcription when its C-terminal region is deleted (Chen *et al*, 1995), although of course this effect may also be due to a cell-type or promoter specific effect as observed with TTF-1.

The study of *Drosophila clift* provides biological data to support XBap acting as a transcriptional repressor rather than an activator. In *bagpipe* null mutants the domain of *clift* expression is expanded in a manner consistent with extra SGP cells being recruited from cells that would normally become visceral mesoderm. This would suggest that wild-type *bagpipe* function may cause the repression SGP cell development (Boyle *et al*, 1997). Although there is no evidence that the repression of clift is a direct effect of *bagpipe* the demonstration that a *bagpipe* relative can exhibit repressive activity may strengthen this possibility. Unfortunately there is, as yet, no vertebrate homologue of *clift*.

6.4 Future work

The first experiment would be to test a construct containing only the homeodomain and the C-terminus of XBap as a GAL4DNB fusion in the Gal4/SV40 assay. This would determine whether the C-terminus of XBap contains an activator domain in addition to the N-terminal repressor domain, as observed in TTF-1 and Nkx-2.5. This assay would be improved by cloning XBap DNA binding sites upstream of the SV40/firefly luciferase reporter in place of the Gal4 DNA binding sites. This would remove the need for Gal4DNB fusion's allowing wild-type XBap protein to be studied directly. A panel of deletions of XBap could then be studied in order to map the position of any activator or repressor domains within the protein.

In view of the study by De Felice *et al*, 1995, in which the transcriptional repression of the TN containing region of TTF-1 was effected by both the promoter and cell context it would be interesting to repeat this assay in a different cell line. The issue remains, of course, that these cells are mammalian not amphibian. It may, therefore, be more informative to carry out studies of the mouse NK3 relatives using this system in a range of cell lines. The cell lines TM3 or TM4 may be good candidates considering the expression of Nkx-3.1 in testis.

Chapter 7

Final Discussion

7.0 Chapter 7: Final Discussion

7.1 Introduction

The aim of this project was first to isolate a *Xenopus* relative of *bagpipe*. This aim was successfully achieved by screening a *Xenopus* stage 32-36 cDNA library. Once the clone had been fully sequenced binding site selection was used to determine the DNA binding specificity of XBap. The site identified proved to be a novel homeoprotein DNA binding site which led to a series of experiments to determine the DNA binding characteristics of XBap, *Drosophila bagpipe* and the newly identified murine relatives. The thesis was completed with an investigation into the transcriptional regulatory abilities of XBap using a number of different systems. The implications of the results have been extensively discussed within each chapter, however, there are issues for discussion which are of relevance to all aspects of the thesis.

7.2 Interacting factors

The potential for interaction with other proteins in a cell type or promoter specific manner could influence the *in vivo* function of XBap by altering its DNA binding specificity and affinity for key target genes as well modifying its ability to act as a transcriptional activator or repressor.

The possibility that the DNA binding specificity of XBap is dependent on the formation of homodimers is an issue that could not be resolved by the experiments described here. Certain regions of the XBap protein are implicated to be involved in such interactions by work conducted on related proteins for example, sequences C-terminal to the homeodomain have been shown to be important for heterodimer formation of C. .legans MEC-3 and the yeast MATα2 proteins (Mann *et al*, 1995). The crystal structure of eve, which is particularly interesting as it has also been shown to bind to a paired site, has revealed that the N-terminal arm, and the polypeptide that continues from it, are well positioned to interact with the adjacent HD (Hirsch *et al*, 1995). The N-terminal region of the homeodomain has also been shown to be critical to protein-protein interactions of

Ubx and Exd and for HOX interactions (Chang *et al*, 1996, Zappavigna *et al*, 1994). Interestingly, alignment of all NK class genes (Fig. 7) shows that the NK3 class genes are most distinguishable from their NK2 class relatives by the sequence of their N-terminal arm. In view of these studies it would be pertinent to generate a panel of deletion constructs of the full XBap protein and observe the effect on DNA binding in EMSA assays and on complex formation when co-translated. The construction of chimaeric proteins between XBap and TTF-1, swapping the N-terminal arms and sequences C-terminal to the homeodomain, may also be revealing.

Previous studies have also produced examples in which transcriptional repression characteristics of homeodomain proteins are modified by interaction with other proteins. As mentioned earlier the ability of *Drosophila* Goosecoid to strongly repress the activation of GR by Orthodenticle relies on specific protein-protein interactions. This effect has been shown to be dependent on both the eh-1 domain, which is homologous to the NK class TN domains, and the N-terminal arm/helix 1 region of the homeodomain (Mailhos *et al*, 1998). The studies of TTF-1 deletion constructs discussed in chapter five revealed that the transcriptional behaviour of the eh-1/TN containing domain appears to be dependent on both the cell and promoter context. A potential mechanism for this context specific activity may be co-operation with other factors.

Considering the potential implications of proteins able to interact with XBap an extremely rewarding path for future research would be to conduct a yeast-2-hydrid assay (Fields and Sternglanz, 1994). Initial experiments would use XBap as the 'bait' and a *Xenopus* cDNA library as the test in an attempt to identify any factors able to interact with XBap, possibly including XBap itself. Experiments could then be carried out to determine the expression pattern of any interacting factors and compare it with XBap expression. The effect of any interacting factors on the DNA binding characteristics of XBap could also be studied using EMSA assays.

7.3 Potential biological targets

Much could be revealed both about the DNA binding and transcriptional abilities of XBap by the identification of a biological target of XBap. In *Drosophila*, null mutants of the gene *Drosophila bap* have suggested that *clift* may be a target for down regulation by *bagpipe*. In view of the observation that a C-terminal deletion of XBap can have transcriptional repression activities it would very rewarding to study the effect of XBap on any *clift* homologs. The wild-type expression pattern of *Drosophila vimar* and the loss of *vimar* expression in the prospective hindgut region of *bagpipe* mutants has also placed *vimar* downstream of *bagpipe*. Although there is as yet no evidence that either *clift* or *vimar* are direct targets of *bagpipe* it would be interesting to screen their promoters for potential Bagpipe DNA binding sites. The identification of *in vivo* Bagpipe sites may answer key questions as to the DNA binding specificity of NK3 class proteins, particularly whether they discriminate between 5'-CAAG-3' and 5'-TAAG-3' cores.

The divergence in expression pattern of the vertebrate *NK3* class genes, with *XBap* expression in the facial cartilage, *Bapx1* expression in the presclerotome and cartilage and *Nkx-3.1* expression in the male urogenital system may suggest that vertebrate *NK3* class genes have evolved new functions. However, it would still be of interest to clone a vertebrate relatives of *clift* and *vimar* as the two proteins may have co-evolved with *Xenopus bagpipe*.

Further information as to the biological function of *XBap* in *Xenopus* could also be gained from over-expression studies. Although these studies are inherently difficult to interpret, if XBap normally acts as a repressor it may be possible to identify certain tissue types, which are inhibited by the ectopic expression of *XBap*. A more simple initial approach would be to study the phenotype of murine knockouts of the *Bapx1*, such studies are most likely in progress.

7.4 Control of *XBap* expression

In *Drosophila tinman* mutants, *bagpipe* expression is not activated in the dorsal mesoderm suggesting that *bagpipe* may be a target of *tinman*. Considering the expression of *XBap* in regions of the developing gut muscle, along with it's apparently diverged expression in the cartilage of the jaw, it would be interesting to determine whether *XBap* control has been conserved. The most direct approach to addressing this issue would be to ectopically express X*Nkx-2.5* and X*Nkx-2.3* in *Xenopus* embryos to see if they can induce expression of *XBap*. In addition, isolation of the *XBap* promoter in order to verify whether or not it contains XNkx-2.3 or XNkx-2.5 binding sites could be extremely rewarding. The *XBap* promoter could potentially be isolated by screening a genomic library with the 3'untranslated region of *XBap*.

7.5 Additional *Xenopus* genes related to *Drosophila bagpipe*

Considering the presence of more than one *NK3* gene in mouse, humans and *planaria*, as well as the family of genes that appear to represent *tinman* in all vertebrates to date, it seemed likely that *Xenopus* would prove to have more than one *NK3* class gene. Recently, a second *bagpipe*-related gene, *zampogna* (*zax*) was isolated by screening a *Xenopus* tadpole stage cDNA library with the *XBap* homeobox (Newman *et al*, 1999). The *zax* cDNA encodes a protein containing both TN and NK2 domains and the homeodomain is 93% identical to *XBap* and 83% identical to *bap*. The sequence identity of *XBap* and *zax* over the entire coding region is relatively low at approximately 58%, indicating that these two sequences represent distinct genes rather than being closely related copies often present in the pseudotetraploid *Xenopus Laevis* genome (Newman *et al*, 1999).

Zax expression is first detected by whole mount in situ hybridisation at the early tailbud stage (stage 30), as a set of vertical stripes in the region of the pharangeal arches. A second domain of expression in the head is found in the position of the future infrarostral cartilage, a component of the lower jaw. *XBap* is also epxressed in a subset of facial cartilages, but is absent from the infrarostral cartilage. Beginning at the mid tailbud stage,

zax is also expressed in the muscular layer of the midgut at the level of the future stomach, this domain of expression is similar to that of *XBap*. These studies have therefore revealed complementary and overlapping domains of expression of *XBap* and zax, consistent with the process of gene duplication and divergence (Newman *et al*, 1999). It would be interesting to study both the DNA binding characteristics and transcriptional abilities of zax in the same assays as employed here with *XBap*.

7.6 Conclusion

The experiments conducted throughout this thesis have resulted in the isolation of a *Xenopus* relative of *Drosophila bagpipe*. The DNA binding specificity of XBap has been determined and compared with *Drosophila* Bagpipe and the murine NK3 class proteins, and finally the ability of XBap to act as transcriptional repressor in one context has been demonstrated. The questions arising from this research have been discussed and the possible paths to answering them have been explored in the conclusion of this thesis.

8.0 References

Aziapzu, N., and Frasch, M. (1993). *tinman* and *bagpipe*: two homeobox genes that determine cell fates in the dorsal mesoderm of *Drosophila*. *Genes and Development* **7,** 1325-1340.

Azpiazu, N., Lawrence, P., Vincent, J.-P., and Frasch, M. (1996). Segmentation and specification of the *Drosophila* mesoderm. *Genes and Development* **10,** 3183-3194.

Balavoine, G. (1996). Identification of members of several homeobox genes in a *planarian* using ligation mediated polymerase reaction technique. *Nucleic Acids Research* **24,** 1547-1553.

Bi, W., Wu, L., Coustry, F., Combruggre, B. D., and Mauty, S. (1997). DNA binding specificity of the CCAAT-binding factor CBF/NF-Y. *Journal of Biological Chemistry* **272,** 26562-26572.

Bieberich, C. J., Fujita, K., He, W., and Jay, G. (1996). Prostate specific and androgen-dependant expression of a novel homeobox gene. *Journal of Biological Chemistry* **271,** 31779-31782.

Bober, E., Baum, C., Braun, T., and Arnold, H. (1994). A novel NK-related mouse homeobox gene: expression in central and peripheral nervous structures during embryonic development. *Developmental Biology* **162,** 288-303.

Bodmer, R. (1993). The gene *tinman* is required for specification of the heart and visceral muscles in *Drosophila*. *Development* **118,** 719-729.

Boyle, M., Bonini, N., and Nardo, S. D. (1997). Expression and function of *clift* in the development of somatic gonadal precursors within the *Drososphila* mesoderm. *Development* **124,** 971-982.

Brand, T., Andree, B., Schneider, A., Buchberger, A., and Arnold, H. (1997). Chicken *Nkx-2.8*, a novel homeobox gene expressed during early heart and foregut development. *Mechanisms of Development* **64,** 53-59.

Buchberger, A., Pabst, O., Brand, T., Seidl, S., and Arnold, H. (1996). Chick *Nkx-2.3* represents a novel family member of vertebrate homologues to the *Drosophila* homeobox gene *tinman*: differential expression of c*Nkx-2.3* and c*Nkx-2.5* during heart and gut development.

Burglin, T. (1994). "The homeobox guidebook." Oxford University Press, New York.

Castelein, H., Derclerq, P. E., and Boles, M. (1997). DNA binding preferences of PPARX/RXR∝ heterodimers. *Biochemica and Biophysica Research Communications* **233,** 91-95.

Cavener, D. R., and Ray, S. C. (1991). Eukaryotic start and stop translation sites. *Nucleic Acids Research* **19,** 3185-3192.

Chambers, A. E., Logan, M., Kotecha, S., Towers, N., Sparrow, D., and Mohun, T. J. (1994). The RSRF/MEF2 protein SL1 regulates cardiac muscle-specific transcription of a myosin light-chain gene in *Xenopus* embryos. *Genes and Development* **8,** 1324-1334.

Chan, S., Ryoo, H., Gould, A., Krumlauf, R., and Mann, R. S. (1997). Switching the *in vivo* specificity of a minimal Hox-responsive element. *Development* **124,** 2007-2014.

Chan, S.-K., Popperl, H., Krumlauf, R., and Mann, R. S. (1996).

An extradenticle-induced conformational change in a HOX protein overcomes an inhibitory function of the conserved hexapeptide motif. *The EMBO Journal* **15,** 2476-2487.

Chang, C.-P., Brochiori, L., Shen, W.-f., Langman, C., and Cleary, M. L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specifities across the Hox locus. *Molecular and Cellular Biology* **16,** 1734-1754.

Chen, Y. C., and Schwartz, J. (1995). Identification of novel DNA binding targets and regulatory domains of a murine *tinman* homeodomain factor, *Nkx-2.5*. *The Journal of Biological Chemistry* **270,** 15628-15633.

Damante, G., D.Fabbro, Pellizzari, L., Civitareale, D., Guazzi, S., Polycarpou-Schwartz, M., Cauci, S., Quadrifoglio, F., Formisano, S., and Lauro, R. D. (1994). Sequence-specific DNA recognition by the thyroid transcription factor-1 homeodomain. *Nucleic Acids Research* **22,** 3075-3083.

Damante, G., and Lauro, R. D. (1991). Several regions of Antennapedia and thyroid transcription factor 1 homeodomains contribute to DNA binding specificity. *Proceedings of the National Academy of Science, USA* **88,** 5382-5392.

Damante, G., Pelizzari, L., Esposito, G., Fogolari, F., Viglino, P., Fabbro, D., Tell, G., Formisano, S., and Lauro, R. D. (1996). A molecular code dictates sequence specific DNA recognition by homeodomains. *The EMBO Journal* **15,** 4992-5000.

Del Sal, G., Manfiolletti, G., and Schneider, C. (1988). A one-tube plasmid mini-preparation suitable for sequencing. *Nucleic acids Research* **16,** 1220.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12,** 387-395.

Dijk, V., and Murre, C. (1994). Extradenticle raises the DNA binding specificity of homeotic selector gene products. *Cell* **78,** 617-624.

Dohrmann, C., Azpiazu, N., and Frasch, M. (1990). A new *Drosophila* homeobox gene is expressed in mesodermal precursor cells of distinct muscles during embryogenesis. *Genes and Development* **4,** 2098-2111.

Driever, W., and Nusslein-Volhard, C. (1989). The bicoid protein is a positive regulator of hunchback transcription in the early *Drosophila* embryo. *Nature* **337,** 138-143.

Evans, S. M., Yan, W., Murillo, M. P., Ponce, J., and Papalopulu, N. (1995). *tinman*, a *Drosophila* homeobox gene required for heart and visceral mesoderm specification, may be represented by a family of genes in vertebrates: *XNkx-2.3*, a second vertebrate homolog of *tinman*. *Development* **121,** 3150-3161.

Fabbro, D., Tell, G., Pellazari, L., Leonardi, A., Pucillo, C., Lonigro, R., and Damante, G. (1995). Definition of the DNA-binding specificity of TTF-1 homeodomain by chromatographic selection of binding sequences. *Biochemical and biophysical research communications* **213,** 781-788.

Felice, M. D., Damante, G., Zannini, M., Francis-Lang, H., and Lauro, R. D. (1995). Redundant domains contribute to transcriptional activity of the thyroid transcription factor. *Journal of Biological Chemistry* **270,** 26649-26656.

Frasch, M. (1995). Induction of visceral and cardiac mesoderm by ectodermal Dpp in the early *Drosophila* embryo. *Nature* **374,** 464-467.

Gajewski, K., Kim, Y., Choi, C. y., and Shulz, R. A. (1998). Combinatorial control of *Drosophila* mef2 gene expression in cardiac and somatic muscle cell lineages. *Developmental Genetics and Evolution* **208,** 382-392.

Gajewski, k., Kim, Y., Lee, Y. M., Olson, E. N., and Schulz, R. A. (1997). D-mef2 is a target for Tinman activation during *Drosophila* heart development. *EMBO Journal* **16,** 515-522.

Garcia-Fernandez, J., Baguna, J., and Salo, E. (1991). *Planaria*n homeobox genes: cloning, sequence analysis and expression. *Proceedings of the National Academy of Science U.S.A.* **88,** 7338-7342.

Gehring, W. J., M.Affolter, and Burglin, T. (1994 A). Homeodomain proteins. *Annual Review Biochemistry* **63,** 487-526.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994 B). Homeodomain-DNA recognition. *Cell* **78,** 211-223.

Gehring, W. J., Muller, M., Affolter, M., Percival-Smith, A., Billeter, M., Qian, Y. Q., Otting, G., and Wuthrich, K. (1990). The structure of the homeodomain and its functional implications. *Trends in Genetics* **6,** 323-329.

Goutte, C., and Johnson, A. D. (1993). Yeast a1 and alpha 2 homeodomain proteins form a DNA-binding activity with properties distinct from those of either protein. *Journal of Molecular Biology* **223,** 359-371.

Guazzi, S., Price, M., De Felice, M., Damante, G., Mattei, M., and Di Lauro, R. (1990). Thyroid nucleur factor (TTF-1) contains a homeodomain and displays a novel DNA binding specificity. *The EMBO Journal* **9,** 3631-3639.

Hamdan, H., Liu, H., Delamos, R., and Minoo, P. (1998). Structure of the human Nkx-2.1 gene. *Biochemica and Biophysica Acta.* **1396,** 336-348.

Hanahan, D. (1985). Techniques for transformation of E-Coli in DNA cloning: a practical appoach. *Oxford: IRL Press Limited* **1,** 109-135.

Hartigan, D. J., and Rubenstein, J. L. R. (1996). The cDNA sequence of murine Nkx-2.2. *Gene* **168,** 271-272.

Harvey, R. P. (1996). *NK-2* Homeobox Genes and Heart Development. *Developmental Biology* **178,** 203-216.

He, W., Sciavolino, P. J., Wing, J., Augustus, M., Hudson, P., Meissner, P. S., Curtis, R. T., Shell, B. K., Bostwick, D. G., Tindall, D. J., Gelmanns, E. P., Abate-Shen, C., and Carter, K. C. (1997). A novel prostate-specific, androgen-regulated homeobox gene (NKX-3.1) that maps to 8p21, a region frequently deleted in prostate cancer. *Genomics* **43,** 69-77.

Hirsch, J. A., and Aggarwal, A. K. (1994). Structure of the Even-skipped homeodomain complexed to AT-rich DNA: new perspectives on homeodomain specificity. *The Embryology Journal.*

Hoey, T., and Levine, M. (1988). Homeobox proteins as sequence -specific transcription factors. *Cell* **55,** 537-540.

Inoue, H., Rudnick, A., German, M. S., Veile, R., Donis-Keller, H., and Permutt, M. A. (1996). Isolation, characterisation and chromosomal mapping of the human Nkx6.1 gene (NKX6A), a new pancreatic islet homeobox gene. *Genomics* **40,** 367-370.

Jimenez, F., Martin-Morris, L., Velasco, L., Chu, H., Sierra, J., R, R. D., and White, K. (1995). vnd, a gene required for early neurogenesis of *Drosophila*, encodes a homeodomain protein. *The EMBO journal* **14,** 3487-3495.

Kalionis, B., and O'Farrell, P. H. (1993). A universal target sequence is bound *in vitro* by diverse homeodomains. *Mechanisms of Development* **43,** 57-70.

Kim, Y., and Nirenberg, M. (1989). *Drosophila* NK-homeobox genes. *Proceedings of the National Academy of Science, USA* **86,** 7716-7720.

Kissenger, C. R., Liu, B., Martin-Bianco, E., Kornberg, T. B., and Pabo, C. O. (1990). Crystal structure of an Engrailed homeodomain-DNA complex at 2.8A resolution: A framework for understanding homeodomain-DNA interactions. *Cell* **63,** 579-590.

Kozak, M. (1987). At least six nucleotides preceeding the AUG initiator codon enhance translation in mammalian cells. *Journal of Molecular Biology* **196,** 947-950.

Krieg, P. A., and Melton, D. A. (1984). Functional messenger RNAs are produced by SP6 *in vitro* transcription of cloned cDNAs. *Nucleic Acids Research* **12,** 7057-7070.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680-685.

Laughon, A. (1991). DNA binding specificity of homeodomains. *Biochemistry* **30,** 11357-11367.

Lazzaro, D., Price, M., De Felice, M., and Di Lauro, R. (1991). The transcription factor TTF-1 is expressed at the onset of thyroid and lung morphogenesis and in restricted regions of the foetal brain. *Development* **113,** 1093-1104.

Lee, K. H., Xu, Q., and Breitbart, R. E. (1996). A new *tinman*-related gene, *nkx-2.7,* anticipates the expression of *nkx-2.5* and *nkx-2.3* in *Zebrafish* heart and pharangeal endoderm. *Developmental Biology* **180,** 722-731.

Lee, Y. M., Parks, T., Schulz, R. A., and Kim, Y. (1997). Twist-mediated activation of the *NK-4* homeobox gene in the visceral mesoderm of *Drosophila* requires two distinct clusters of E-box regulatory elements. *The Journal of Biological Chemistry* **272,** 17531-17541.

Lints, T. J., Parsons, L. M., Hartley, L., Lyons, I., and Harvey, R. P. (1993). *Nkx-2.5*: a novel murine homeobox gene expressed in early heart progenitor cells and their myogenic descendants. *Development* **119,** 419-431.

Lo, P. C., and Frasch, M. (1998). *bagpipe*-dependent expression of vimar, a novel Armadillo-repeats gene, in *Drosophila* visceral mesoderm. *Mechanisms of Development* **72,** 65-75.

Lu, Q., and Kamps, M. P. (1997). Heterodimerization of Hox proteins with Pbx1 and oncoprotein E2a-Pbx1 generates unique DNA-binding specificities at nucleotides predicted to contact the N-terminal arm of the Hox homeodomain - demonstration of Hox-dependant targeting of E2a-Pbx 1 *in vivo*. *Oncogene* **14,** 75-83.

Lyons, I., Parsons, L. M., Hartley, L., Li, R., Andrews, J., and Robb, L. H., Harvey R.P. (1995). Myogenic and morphognetic defects in the heart tubes of murine embryos lacking the homeobox gene *Nkx-2.5*. *Genes and Development* **9,** 1654-1666.

Mailhos, C., Andre, S., Mollereau, B., Goreily, A., Hemmanti-Brivanlou, A., and Desplan, C. (1998). *Drosophila* Goosecoid requires a conserved heptapeptide for repression of paired-class homeoprotein activators. *Development* **125,** 937-947.

Mann, R. S. (1995). The specificity of homeotic gene function. *Bioessays* **17,** 855-863.

McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and Axial Patterning. *Cell* **68,** 283-302.

Muller, M., Affolter, M., Leupin, W., Otting, G., Wuthrich, K., and Gehring, W. J. (1988). Isolation and sequence-specific DNA binding of the Antennapedia homeodomain. *The EMBO Journal* **7,** 4299-4304.

Nardelli-Haefliger, D., and Shankland, M. (1993). *Lox 10*, a member of the *NK-2* homeobox gene class, is expressed in a segmental pattern in the endoderm and in the cephalic nervous system of the leech Helobdella. *Development* **118,** 877-892.

Newman, C. S., Grow, M. W., Cleaver, O., F. Chia, F., and Krieg, P. (1997). *XBap*, a vertebrate gene related to *bagpipe*, is expressed in developing craniofacial structures and in anterior gut muscle. *Developmental Biology* **181,** 223-233.

Nicolas, S., Caubit, C., Massacrier, A., Lau, P., and LePuroo, Y. (1999). Two *Nkx-3*-related genes are expressed in the adult wall and regenerating central nervous system of the urodele Pleurodele Walt1. *Developmental Genetics* **24,** 319-328.

Nirenberg, M., Nakayama, K., Nakayama, N., Kim, Y., Mellerick, D., Wang, L., Webber, K. O., and Lad, R. (1995). The *NK-2* homeobox gene and the early development of the central nervous system of *Drosophila*. *Annals New York academy of science* **758,** 224-242.

Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., Gehring, W., and Wuthrich, K. (1990). Protein-DNA contacts in the structure of a homeodomain-DNA complex determined by nucleur magnetic resonance spectroscopy in solution. *The EMBO Journal* **9,** 3085-3092.

Pellizari, L., Tell, G., Fabbro, D., Pucillo, C., and Damante, G. (1997). Functional interference between contacting amino acids of homeodomains. *FEBS Letters* **407,** 320-324.

Pollock, R., and Treisman, R. (1991 A). Human SRF-related proteins: DNA-binding properties and potential regulatory targets. *Genes and Development* **5,** 2327-2341.

Pollock, R. and Treisman, R. (1991 B). A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Research,* Vol **18,** No. 21, 6197-6204.

Popperl, H., Bienz, M., Studer, M., Chan, S. K., Aparicio, S., Brenner, S., Mann, R. S., and Krumlauf, R. (1995). Segmental expression of *hoxb-1* is controlled by a highly conserved auto-regulatory loop dependant upon Exd/Pbx. *Cell* **81,** 1031-1042.

Prescott, J. L., Blok, L., and Tindall, D. J. (1998). Isolation and androgen regulation of a novel human homeobox gene, Nkx-3.1. *Prostate* **35,** 71-80.

Price, M., Lazzaro, D., Pohl, T., Mattei, M.-G., Ruther, U., Olivo, J.-C., Duboule, D., and Di-Lauro, R. (1992). Regional expression of the homeobox gene *Nkx-2.2* in the developing mammalian forebrain. *Neuron* **8,** 241-255.

Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W., and Wuthrich, K. (1989). The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: Comparison with prokaryotic repressors. *Cell* **59,** 573-580.

Reecy, J. M., Yamada, M., Cummings, K., Sosic, D., Chen, C., Eichele, G., Olson, E. N., and Schwartz, R. J. (1997). Chicken *Nkx-2.8*: A novel homeobox gene expressed in early heart progenitor cells and pharangeal pouch-2 and -3 endoderm. *Developmental Biology* **188,** 295-311.

Riechmann, V., Rehorn, K., Reuter, R., and Leptin, M. (1998). The genetic control of the distinction between fat body and gonadal mesoderm in *Drosophila*. *Development* **125,** 713-723.

Rinkwitz-Brandt, S., Justus, M., Oldenettel, I., Arnold, H., and Bober, E. (1995). Distinct temporal expression of mouse Nkx-5.1 and Nkx-5.2 homeobox genes during brain and ear development. *Mechanisms of Development* **52,** 371-381.

Saha, M. S., Michel, R. B., Gulding, K. M., and Grainger, R. M. (1993). A *Xenopus* homeobox gene defines dorsal-ventral domains in the developing brain. *Development* **118,** 193-202.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual. *Cold Spring Harbour Laboratory Press*.

Sanger, F., Nicklen, S., and Coulsen, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science, USA* **74,** 5463-5467.

Schubert, F. R., Fainsod, A., Gruenbaum, Y., and Gruss, P. (1995). Expression of the novel homeobox gene *Sax-1* in the developing nervous system. *Mechanisms of Development* **51,** 99-114.

Sciavolino, P., Abrams, E. W., Yang, L., Austenberg, L. P., Shen, M. M., and Abate-Shen, C. (1997). Tissue specific expression of murine *Nkx-3.1* in the male urogenital system. *Developmental Dynamics* **209**.

Scultheiss, T. M., Xydas, S., and Lassar, A. B. (1995). Induction of avian cardiac myogenesis by anterior endoderm. *Development* **121,** 4203-4214.

Shanmugam, K., Featherstone, M. S., and Saragori, H. U. (1997). Residues flanking the HOX YPWM motif contribute to co-operative interactions with PBX. *Journal of Biological Chemistry* **272,** 19081-19087.

Smith, S and Jaynes, J (1996). A conserved region of Engrailed, shared among all en-, gsc-, Nk1-, Nk2- and msh-class homeoproteins, mediate active transcriptional repression in vivo. *Development* **122**, 3141-3150.

Staden, R. (1982). Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Research* **10**, 4731-4751.

Staden, R. (1984). Automation of the handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Research* **12**, 521-538.

Staehling-Hampton, K., Hoffmann, F. M., Baylies, M. K., Rushton, E., and Bate, M. (1994). dpp induces mesodermal gene expression in *Drosophila*. *Nature* **372**, 783-467.

Tanaka, M., Chen, Z., Bartunkova, S., Yamasaki, N., and S. Izumo. (1999). The cardiac homeobox gene *Csx/Nkx2.5* lies genetically upstream of multiple genes essential for heart development. *Development* **126**, 1269-1280.

Tonissen, K. F., Drysdale, T. A., Lints, T. J., Harvey, R. H., and Krieg, P. A. (1994). X*Nkx-2.5*, a *Xenopus* gene related to *Nkx-2.5* and *tinman*: Evidence for a conserved Role in cardiac development. *Developmental Biology* **162**, 325-328.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**, 553-562.

Triboli, C., Frasch, M., and Lufkin, T. (1997). *Bapx1*:an evolutionary conserved homologue of the *Drosophila bagpipe* homeobox gene is expressed in the sphlanchnic mesoderm of the embryonic skeleton. *Mechanisms of Development* **65,** 145-162.

Tsao, D., Gruschus, J. M., Wang, L., Nirenberg, M., and Ferretti, J. A. (1995). The three-dimensional solution structure of the NK-2 homeodomain from *Drosophila*. *Journal of Molecular Biology* **251,** 297-307.

Turbay, D., Wechler, S. B., Blanchard, K. M., and Izumo, S. (1996). Molecular cloning, chromosomal mapping, and characterisation of the human cardiac-specific homeobox gene *hCsx*. *Molecular Medicine* **2,** 86-96.

Wilson, D., Sheng, G., Lecuit, T., Dostani, N., and Desplan, C. (1993). Cooperative dimerization of paired class homeodomains on DNA. *Genes and Development* **7,** 2120-2134.

Wilson, D. S., Sheng, G., Jun, S., and Desplan, C. (1996). Conservation and diversification in homeodomain-DNA interactions: A comparitive analysis. *Proc. Natl. Acad. Sci* **93,** 6886-6891.

Yan, Z. H., Medvedev, A., Hirose, T., Gotoh, H., and Jetten, A. M. (1997). Characterisation of the response element and DNA binding properties of the nucleur orphan receptor germ cell nucleur factor/retinoid receptor-related testis associated receptor. *The Journal of Biological Chemistry* **272,** 10565-10572.

Yoshiura, K., and Murray, J. C. (1997). Sequence and chromosomal assignment of human BAPX1 a *bagpipe* related gene, to 4p16.1: A candidate gene for skeletal dysplasia. *Genomics* **45,** 425-428.

Zappavigna, V., Falciola, L., Citterich, M. H., Mavilio, F., and Bianchi, M. E. (1996). HMG1 interacts with HOX proteins and enhances their DNA binding and transcriptional activation. *The EMBO Journal* **15,** 4981-4991.

Zhang, H., Scholl, R., Browse, J., and Somerville, C. (1988). Double-stranded DNA sequencing as a choice for DNA sequencing. *Nucleic Acids Research* **16,** 9878.

Addendum to References

Chen, C. and Schwartz, R. (1995). Identification of novel DNA binding targets and regulatory domains of a murine Tinman homeodomain factor, Nkx-2.5. *The Journal of Biological Chemistry* **270**, 15628-15633.

Damante, G., Tell, G., Formisano, S., Fabbro, D., Pellizzari, L. and Di Lauro, R. (1993). Effect of salt concentration on TTF-1 HD binding to specific and non-specific DNA sequences. *Biochemical and Biophysical Research Communications* **197**, 632-638.

Etkin, L. and Balcells, S. (1985). Transformed *Xenopus* embryos as a transient expression system to analyze gene expression at the mid blastula transition. *Developmental Biology* **108**, 173-178.

Fields, S. and Sternglanz, R. (1994). The two hybrid system:an assay for protein-protein interactions. *Trends in Genetics,* **10** (8) 286-292.

Fu, Y., Yan, W., Mohun, T.J. and Evans, S.M. (1998). Vertebrate *tinman* homologues *XNkx2-3* and *XNkx2-5* are required for heart formation in a functionally redundant manner. *Development* **125**, 4439-4449.

Newman, C. and Krieg, P. (1999). The *Xenopus bagpipe*-related homeobox gene *Zampogna* is expressed in the pharyngeal endoderm and the visceral musculature of the midgut. *Developmental Genetics and Evolution* **209**, 132-134.

Woolberger, C., Vershon, A., Liu, B., Johnson, A. and Pabo, C.O. (1991). Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**, 517-528.